Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Leon Noe Jovan

Priprava podatkov za občinskega virtualnega asistenta s pomočjo strojnega učenja

magistrsko delo

študijski program druge stopnje
računalništvo in informatika

Mentor: izr. prof. dr. Matjaž Kukar
Somentor: prof. dr. Matjaž Gams

Ljubljana, 2016

Izjava o avtorstvu magistrskega dela

Spodaj podpisani Leon Noe Jovan sem avtor magistrskega dela z naslovom:

Priprava podatkov za občinskega virtualnega asistenta s pomočjo strojnega učenja

S svojim podpisom zagotavljam, da:

- sem magistrsko delo izdelal samostojno pod mentorstvom izr. prof. dr. Matjaža Kukarja in somentorstvom prof. dr. Matjaža Gamsa,
- so elektronska oblika magistrskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko magistrskega dela,
- soglašam z javno objavo elektronske oblike magistrskega dela v zbirki "Dela FRI".

V Ljubljani, 20. junija 2016 Podpis avtorja:

Zahvaljujem se mentorju izr. prof. Matjažu Kukarju za vso strokovno pomoč, napotke in ideje pri izdelavi magistrskega dela.

Zahvala gre tudi somentorju prof. Matjažu Gamsu za vso podporo in dostop do vsebin projekta Asistent, ki sem jih potreboval za izdelavo magistrskega dela.

Hvala tudi Tanji Breznik za kvaliteten ročni pregled rezultatov.

Posebna zahvala gre tudi družini in punci Maji, ki so mi stali ob strani vsa leta študija.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
SVM	support vector machine	metoda podpornih vektorjev
MRR	mean reciprocal rank	srednja recipročna uvrstitev
RPC	ranking by pairwise comparison	uvrščanje s parno primerjavo
RAkEL	random k labelsets	naključne množice k oznak
BR	binary relevance	binarna relevanca
LP	label powerset	potenčne množice oznak
priklic@k	recall at k	priklic pri k
IG	information gain	informacijski prispevek

Povzetek

Cilj magistrskega dela je bil razvoj postopka priprave podatkov za občinskega virtualnega asistenta, ki odgovarja na razna postavljena vprašanja v zvezi z dejavnostmi občin v Sloveniji. Namen postopka je nadomestiti ali olajšati ročno pripravo baze znanja virtualnega asistenta. Zato smo preučili in uporabili različna področja strojnega učenja, kot so večznačna klasifikacija, uporaba šibko označenih podatkov, razvrščanje v skupine in besedilno rudarjenje. V magistrskem delu smo predstavili postopek, ki za različna vprašanja v zvezi z dejavnostjo občin poišče najbolj ustrezne spletne strani, ki dajejo odgovore na ta vprašanja. Parametre postopka smo najprej optimizirali na testnih podatkih, nato pa smo ga na podatkih novih občin ovrednotili tudi ročno. Tako smo pridobili resnično oceno kvalitete delovanja razvitega postopka. Rezultati so pokazali, da ta postopek predlaga bolj ustrezne odgovore, kot jih predlaga komercialni spletni iskalnik. Z razvitim postopkom smo torej učinkovito pospešili in poenostavili pripravo podatkov za občinskega virtualnega asistenta. S tem smo olajšali delo pristojnim zaposlenim na občinah, ki so do sedaj ročno vnašali odgovore v bazo znanja občinskega virtualnega asistenta.

Ključne besede: občinski virtualni asistent, večznačna klasifikacija, šibko označeni podatki, razvrščanje v skupine, tekstovno rudarjenje.

Abstract

The main goal of this master's thesis was to develop a procedure that will automate the construction of the knowledge base for a virtual assistant that answers questions about municipalities in Slovenia. The aim of the procedure is to replace or facilitate manual preparation of the virtual assistant's knowledge base. Theoretical backgrounds of different machine learning fields, such as multilabel classification, text mining and learning from weakly labeled data were examined to gain a better understanding of the topic. In this thesis, we present a procedure that finds the most relevant websites to provide answers on various questions relating to the municipality's activities. The procedure's parameters were first optimized using test data, and then the procedure was evaluated manually using data of new municipalities. In this way, we acquired real estimation of the quality of the implemented procedure. The results show that the procedure recommends more relevant answers in comparison to a commercial search engine. The developed procedure therefore effectively speeds up and simplifies data preparation for the municipal virtual assistant. In this way, we facilitate the work of municipality staff who until now had to insert answers into the municipal virtual assistant's knowledge base manually.

Keywords: municipal virtual assistant, multi-label classification, weakly labeled data, clustering, text mining.

Poglavje 1

Uvod

Svetovni splet je ogromna zbirka podatkov, sestavljena iz milijard spletnih strani, ki vsebujejo različne informacije in so medsebojno povezane. Ti podatki so relativno slabo strukturirani in pogosto nepreverjeni. Iskanje informacij na svetovnem spletu je zato pogosto zelo težavno. Splošni spletni iskalniki do neke mere olajšajo iskanje in uporabo, kljub temu pa je iskanje pogosto tudi s pomočjo tovrstnih orodij neuspešno, predvsem pri zahtevnejših in bolj specifičnih poizvedbah. Nizka uspešnost iskanja želenih informacij predstavlja problem tako za uporabnike kot tudi lastnike spletnih strani.

V Sloveniji je na problem učinkovitega iskanja informacij opozorilo tudi Združenje občin Slovenije. Spletne strani občin so pogosto velike, nekatere so sestavljene tudi iz več kot deset tisoč podstrani. Zaradi nerednega posodabljanja so te strani zastarele in slabo organizirane. To predstavlja težavo celo bolj izkušenim uporabnikom interneta, še bolj pa starejšim, hendikepiranim in neveščim uporabnikom.

Predlagana rešitev za ta problem je vzpostavitev inteligentnega virtualnega pomočnika, nameščenega na spletni strani občine, ki odgovarja na vprašanja, postavljena v naravnem jeziku in je sposoben poiskati relevantne odgovore, do neke mere podobno kot človek. V ta namen je nastal projekt Asistent [1], inteligentni vmesnik, ki uporabnikom na enostaven način pomaga do različnih informacij. Prijavitelj projekta je Združenje občin Slo-

venije v sodelovanju z Institutom "Jožef Stefan", Zvezo društev upokojencev Slovenije ter občino Pivka, občino Litija, občino Slovenj Gradec in občino Dobrova-Polhov Gradec.

1.1 Projekt Asistent

Občinski asistent [1] je inteligentni virtualni pomočnik, ki je bil razvit na Institutu "Jožef Stefan". Razume vprašanja v naravnem jeziku in skuša uporabniku ponuditi najboljši odgovor. Prva vzorčna aplikacija občinskega asistenta je namenjena obiskovalcem spletnih stran občin. Z enostavnim uporabniškim vmesnikom, iskanjem in odgovarjanjem v naravnem jeziku občinski asistent pomaga tudi računalniško neveščim uporabnikom spleta, da pridejo do želenih informacij.

Njegovo delovanje temelji na vzpostavitvi brezplačne storitve v oblaku za izdelavo in urejanje prilagojenega virtualnega pomočnika, ki ga bodo občine namestile na svoje spletne strani.

Uporaba občinskega asistenta je preprosta. V vnosno polje vnesemo vprašanje v naravnem jeziku, občinski asistent pa nam odgovori s kratkim odgovorom v oknu pod vnosnim poljem, v ozadju pa se odpre tudi ustrezna spletna stran, na kateri najdemo dodatne informacije. Primer uporabe prikazuje slika 1.1.

Občinski asistent zna poleg splošnega pogovarjanja odgovoriti na približno 500 vprašanj o občini in občinskih storitvah, iz devetih kategorij, podrobneje opisanih v prilogi A.

Predpostavili smo, da je nabor standardnih vprašanj skupen vsem občinam, vendar pa so nekateri odgovori za posamezno občino specifični, zato uporaba virtualnega pomočnika zahteva ločeno pripravo podatkov za vsako občino. Ena od možnosti za uspešno delovanje občinskega asistenta je ročni vnos vseh odgovorov na nabor 500 vprašanj, ki bi ga izvedli uredniki občinskega asistenta na posamezni občini. Ročni vnos odgovorov in pripadajočih povezav spletnih strani sta se izkazala kot zamudno opravilo, ki pripelje do

Slika 1.1: Primer uporabe občinskega virtualnega asistenta.

možnosti napak in do možne neažurnosti podatkov.

Druga možnost je avtomatizacija priprave podatkov za posamezno občino, ki urednika občinskega asistenta razbremeni rutinskega dela, ob tem pa zagotavlja tudi ažurnost podatkov. Razvoj takšnega postopka je bila moja naloga na Institutu "Jožef Stefan". Celoten postopek je sestavljen iz dveh faz avtomatizacije. Prva faza avtomatizacije, ki obsega pripravo kratkih odgovorov na vprašanja uporabnikov občinskih spletnih strani, je bila izvedena že pred časom, zato ni vključena v magistrsko delo [2]. V magistrskem delu bomo predstavili avtomatizacijo iskanja vprašanju ustreznih spletnih povezav, s katero se bo zaključil celoten postopek priprave podatkov.

1.2 Sorodna dela

Vsebina magistrskega dela posega v različna področja strojnega učenja, kot so večznačna klasifikacija (*angl. multi-label classification*), besedilno rudarjenje (*angl. text mining*) in uporaba šibko označenih podatkov (*angl. weakly labeled data*).

V grobem lahko naš problem predstavimo kot primer problema večznačne klasifikacije. Na posamezni spletni strani se lahko nahajajo podatki, ki od-

govorijo na več različnih vprašanj. Ker smo vprašanja predstavili kot razrede, to pomeni, da lahko posamezna spletna stran pripada več kot samo enemu razredu. Poleg metode binarna relevanca, ki je ena od osnovnih metod za večznačno klasifikacijo, smo preizkusili tudi novejšo ansambelsko metodo RAkEL, ki obljublja dobre rezultate [3].

Zaradi določenih posebnosti in omejitev celotnega projekta ter majhne in pomanjkljive množice označenih testnih podatkov klasične mere za ocenjevanje večznačne klasifikacije niso primerne. Zato smo uporabili mere za ocenjevanje uspešnosti, ki se pogosto uporabljajo na področju sistemov za priklic informacij (*angl. information retrieval*) [4, 5].

Zaradi majhnega števila ročno označenih podatkov smo se v magistrskem delu ukvarjali s problemom avtomatskega pridobivanja novih učnih podatkov. Ogromno količino novih podatkov lahko v današnjem času dobimo na spletu, na primer s pomočjo spletnega iskalnika. Vendar pa oznake takšnih podatkov niso zanesljive in lahko vsebujejo veliko šuma. Takšnim podatkom pogosto pravimo šibko označeni podatki (*angl. Weakly labeled data*) [6]. Na področju uporabe šibko označenih podatkov se uporablja več različnih pristopov. Zelo pogost pristop je uporaba delno nadzorovanega učenja, s katerim z uporabo majhne množice označenih podatkov poskusimo označiti šibko označene ali celo neoznačene podatke [7]. Drugi pristop je uporaba robustnih metod [8], ki so bolj odporne na šum v oznakah. Tretji pogost pristop je izboljšanje oznak z uporabo nenadzorovanih metod, na primer razvrščanja dokumentov v skupine in nato rezanja teh skupin [9].

1.3 Cilji

Cilj magistrske naloge je razvoj in ovrednotenje postopka iskanja določenemu vprašanju ustreznih spletnih povezav. S temi spletnimi stranmi bo postopek napolnil bazo občinskega asistenta za posamezno občino. Za vsako iz nabora vnaprej pripravljenih 500 vprašanj je potrebno poiskati najbolj ustrezno spletno stran in jo uvrstiti v podatkovno bazo občinskega asistenta, ki jo nato prikaže uporabniku. Ustrezen začetni postopek iskanja je predpogoj za oblikovanje pravilnih kratkih odgovorov, ki so povzetek vsebine spletne strani. Zato smo preverili možnost uporabe različnih metod večznačne klasifikacije, besedilnega rudarjenja ter uporabe šibko označenih podatkov, ovrednotili učinkovitost teh metod in izbrane vključili v postopek avtomatizacije priprave podatkov.

1.4 Struktura magistrskega dela

Magistrsko delo je razdeljeno na šest poglavij. Po uvodu so v drugem poglavju predstavljene uporabljene množice podatkov ter njihovo pridobivanje. V tretjem poglavju so opisane uporabljene metode s področja besedilnega rudarjenja, odkrivanja skupin in enoznačne in večznačne klasifikacije, razdeljeno na več delov glede na področja. Četrto poglavje opisuje celoten postopek iskanja ustreznih spletnih strani za določeno vprašanje, ki smo ga v delu za magistrsko nalogo razvili. Peto poglavje opisuje uporabljeno testno metodologijo, nato so predstavljeni rezultati testiranja celotnega postopka, izbira parametrov postopka in njihova interpretacija. Na koncu so v šestem poglavju predstavljene sklepne ugotovitve in predlogi za nadaljnje raziskovanje in izboljšavo predlaganega postopka.

Poglavje 2

Opis podatkov

Občinski asistent vsake posamezne občine ima v oblaku lastno podatkovno bazo s shranjenimi odgovori in povezavami do ustreznih spletnih strani. Podatkovna baza je na začetku prazna, napolnimo jo lahko ročno ali pa uporabimo avtomatiziran postopek.

Naša naloga je, da za vsako od 500 vprašanj poiščemo ustrezno spletno stran z relevantnimi informacijami, ki odgovorijo na vprašanje. Te strani se lahko nahajajo praktično kjerkoli na slovenskem spletu. Velik del se sicer nahaja na spletnih straneh občin, vendar se ne moremo omejiti le na te strani, saj spletne strani občin pogosto ne vsebujejo vseh informacij, ki jih potrebujemo. Na primer, občine imajo pogosto za turizem namenjene posebne spletne strani, podobno je tudi z muzeji, šolami ali gledališči. Vprašanje je, katere spletne strani so sploh lahko relevantne v našem delu.

Poglavje opisuje uporabljene množice podatkov in njihovo pridobivanje. Najprej predstavi ročno označene podatke, ki smo jih uporabili za učenje klasifikacijskih modelov. Zaradi majhnega števila teh podatkov smo s pomočjo spletnega iskalnika pridobili še eno množico šibko označenih podatkov. Ti podatki nam služijo v dva namena. Prvič, s temi podatki lahko razširimo učno množico in tako izdelamo boljše klasifikacijske modele. Drugič, s to množico rešujemo tudi problem prostora preiskovanja, omenjen v prejšnjem odstavku. S tem smo učinkovito omejili prostor spletnih strani, ki jih proučujemo.

2.1 Označeni podatki

Označene podatke smo pridobili iz podatkovnih baz občinskega virtualnega asistenta, katerih naslovi spletnih strani so bili vneseni ročno v 11-ih občinah. Označeni podatki so bili osnova za učenje različnih vzorcev, ki nam pomagajo prepoznati nove ustrezne spletne strani.

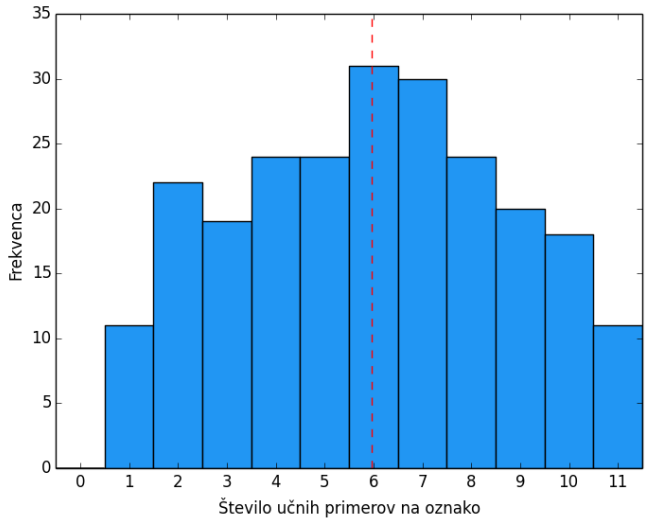
Po ročnem pregledu vseh 500-ih vprašanj smo vprašanja razdelili v dve osnovni skupini - enostavna in kompleksna. Odgovore na podmnožico 265 vprašanj iz prve skupine lahko dobimo na bolj enostaven način, ki je opisan v nadaljevanju (razdelek 4.1). Za odgovore za 235 vprašanj (seznam v prilogi C) iz druge skupine pa smo ocenili, da bomo za iskanje relevantne spletne strani morali uporabiti metode strojnega učenja, saj za njih ne obstaja enostaven način iskanja relevantne spletne strani. Označene učne podatke torej potrebujemo le za 235 različnih vprašanj. Glede na to, da bomo za klasifikacijo spletnih strani uporabili metode večznačne klasifikacije, smo vsako vprašanje predstavili z eno oznako. Torej imamo problem večznačne klasifikacije, kjer nastopa 235 različnih oznak, kjer vsaka oznaka predstavlja določeno vprašanje.

Tabela 2.1 prikazuje osnovne lastnosti označenih podatkov. Dejstvo je, da občine niso vnesle odgovorov na vsa vprašanja, v povprečju je vsaka vnesla povprečno 127 odgovorov na 235 vprašanj, torej dobro polovico. Število vseh odgovorov 11-ih občin, ki opredelijo povezave na ustrezne spletne strani, je 1396, vendar je unikatnih le 831 spletnih strani. Razlog je v tem, da lahko ena spletna stran vsebuje odgovore za več različnih vprašanj. V povprečju vsaka stran odgovarja na 1.68 vprašanja. Za vsako posamezno vprašanje smo dobili v povprečju 5.94 označenih spletnih strani.

Slika 2.1 prikazuje porazdelitev označenih učnih primerov, ki jih imamo za vsako oznako. Iz slike je razvidno, da imamo v povprečju na voljo 5.94 označenih spletnih strani za učenje za posamezno oznako. Za vsako vprašanje imamo na voljo vsaj en označen primer. Za nezanemarljiv del oznak imamo na voljo zelo malo učnih podatkov, v nekaterih primerih samo enega. Največje število primerov za posamezno oznako je 11.

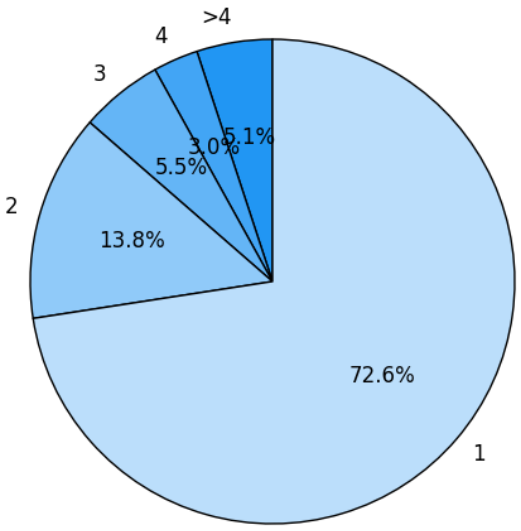
Tabela 2.1: Lastnosti ročno označenih podatkov

atribut	vrednost
število vnosov	1396
število občin	11
število oznak	235
povprečno število vnosov na občino	126.9
število unikatnih povezav	831
povprečno število primerov na oznako	5.94
kardinalnost oznak	1.68



Slika 2.1: Porazdelitev označenih učnih primerov, ki jih imamo za vsako vprašanje. Stolpec višine pri oznaki 6 pomeni, da je 30 vprašanj takih, da je nanj odgovorilo 6 občin.

Slika 2.2 prikazuje relativno frekvenco spletnih strani, glede na to, koliko oznakam pripadajo, oziroma na koliko vprašanj odgovorijo. V povprečju posamezna stran pripada le 1.68 različnim oznakam. Opazimo, da velik del spletnih strani odgovori le na eno vprašanje. Posamezna spletna stran odgovori na največ 22 vprašanj.



Slika 2.2: Slika prikazuje razmerje med številom spletnih strani glede na število pripadajočih oznak oziroma na koliko vprašanj odgovorijo.

Iz osnovne analize podatkov lahko sklepamo, da imamo dejansko na voljo zelo majhno število označenih podatkov. Začetni poskusi so pokazali, da s strojnim učenjem s temi podatki dobimo slabe rezultate, le približno 25% predlaganih strani je relevantnih [10]. Z nekaj primeri za vsako oznako težko zgradimo dober klasifikacijski model, zato stremimo po povečanju učne množice z novimi primeri. To smo dosegli z uporabo šibko označenih podatkov, ki smo jih pridobili s pomočjo spletnega iskalnika.

2.2 Šibko označeni podatki

Zaradi majhnega števila razpoložljivih označenih podatkov/učnih primerov je učenje klasifikacijskih modelov precej oteženo. To velja predvsem v primerih, ko imamo za posamezni razred le nekaj označenih učnih primerov. Ker je ročno označevanje podatkov časovno zelo potratno, potrebujemo avtomatizirano rešitev, ki nam bo našla čim več dodatnih učnih primerov.

Predlagana rešitev je, da s pomočjo iskanja s ključnimi besedami na spletnih iskalnikih pridobimo večje število dodatnih primerov. Takšnim podatkom, kjer so oznake ali pa sami primeri nepopolni, na splošno pravimo šibko označeni podatki [11]. Obstaja sicer več vrst šibko označenih podatkov, v našem primeru gre predvsem za šum v oznakah podatkov. Spletni iskalniki nam na podlagi ključnih besed pogosto ne dajo ustreznih podatkov, zato jih po kvaliteti ne moremo primerjati z ročno označenimi podatki [6]. Na primer, rezultat spletnega iskalnika, ki se je pojavil na enem od zadnjih mest, mogoče niti ni več relevanten za določeno poizvedbo. Šum v oznakah lahko močno zmanjša kvaliteto napovedi [12]. Zato moramo te podatke ustrezno prečistiti in jih prilagoditi za vključitev v učno množico.

Dodatne učne primere smo pridobili s pomočjo spletnih iskalnikov, kjer smo razširili iskanje po širšem naboru občin. Vsako vprašanje smo predstavili z ročno izbranimi ključnimi besedami. Za določitev občine smo k poizvedbi dodali tudi ključno besedo "občina" ter ime občine. Postopek pridobivanja dodatnih učnih primerov opisuje algoritem 1.

Algoritem 1: Psevdokoda za pridobivanje šibko označenih podatkov s pomočjo spletnega iskalnika

```
Data: seznam ključnih besed seznamKljucnihBesed, seznam občin
      seznamObcin
for imeObcine in seznamObcin do
    for kljucneBesede in seznamKljucnihBesed do
        poizvedba ← "občina "+ imeObcine + " " + kljucneBesede;
        rezultati[] ← preberiRezultateSpletnegaIskalnika(poizvedba);
        for rank ← 0 to rezultati.length do
            url ← rezultati[rank];
            shraniRezultat(url,rank,kljucneBesede,imeObcine);
        end
    end
end
```

Z uporabo algoritma 1 smo pridobili nove podatke za dodatnih 85 občin, za vsako občino 235 strani rezultatov, kjer vsaka stran vsebuje približno 10 spletnih povezav za določeno poizvedbo. Skupno smo pridobili 188,113 novih naslovov spletnih strani. Vsaki novi spletni strani smo pripisali, kateri občini pripada, na katero vprašanje potencialno odgovarja, in na kateri poziciji se nahaja na spletnem iskalniku.

Poglavje 3

Metode

To poglavje opisuje metode, ki smo jih uporabili v magistrskem delu. Predstavljeni so pristopi za predstavitev besedilnih dokumentov, razvrščanja dokumentov v skupine in metode za enoznačno in večznačno klasifikacijo. Opisane so tudi mere uspešnosti, s katerimi smo vrednotili delovanje razvitega postopka za pripravo podatkov občinskega asistenta.

3.1 Predstavitev besedilnih dokumentov

Večina klasičnih metod strojnega učenja ne zna delati s čistim besedilom, zato ga moramo najprej predstaviti v ustrezni obliki, da iz besedila pridobimo čim več uporabnih informacij. Običajno želimo besedila predstaviti kot vektorje atributov in njihovih vrednosti.

Takšna predstavitev se pogosto imenuje model vektorskega prostora (*angl. Vector-Space-Model*)/vreča besed (*angl. Bag of words*), kjer je posamezen dokument predstavljen z vektorjem. Elementi vektorja običajno predstavljajo besedo, včasih pa tudi zaporedje sosednjih besed ali črk. Za določanje vrednosti atributov poznamo več različnih tehnik. Najbolj enostavna je binarna, kjer vrednost 1 predstavlja prisotnost besede v dokumentu, 0 pa odsotnost te besede. Drug način je frekvenca besed, ki nam pove, kolikokrat se beseda pojavi v dokumentu. Število pojavitev besed v dokumentu je odvisna

tudi od dolžine dokumenta. Da ta vpliv izničimo, namesto števila pojavitev besed uporabljamo relativno frekvenco, to je verjetnost, da bi pri naključnem izboru besede iz dokumenta izbrali določeno besedo. Ena od pogosto uporabljenih metod je *tf-idf* (*angl. term frequency–inverse document frequency*), s katero dosežemo, da se izrazom, ki se zelo pogosto pojavljajo v korpusu, določi nižja vrednost. *tf-idf* izračunamo s spodnjo enačbo:

$$\begin{aligned} \textit{tf-idf}(t,d) &= \textit{tf}(t,d) * \textit{idf}(t) \\ \textit{tf}(t,d) &= \frac{|t \in d|}{|d|} \\ \textit{idf}(t) &= log\frac{|D|}{|d \in D : t \in d|}, \end{aligned} \tag{3.1}$$

kjer je $|D|$ število vseh elementov v korpusu dokumentov D, število $|d \in D : t \in d|$ pa je število dokumentov, ki vsebujejo besedo t.

Predstavitve dokumentov z opisanimi metodami zanemarjajo dejstvo, da je pomen besed mnogokrat odvisen od konteksta. Ista beseda ima lahko različne pomene, ali pa imajo isti pomen lahko različne besede. Prav tako nam taka predstavitev ne bo razrešila problema podpomenk in drugačnih povezav med različnimi izrazi. Zanemari se tudi vrstni red besed in struktura besedila. Dokument je torej le seznam besed v poljubnem vrstnem redu.

Kljub naštetim pomanjkljivostim pa nam takšni preprosti modeli brez uporabe kakršnekoli informacije o semantiki omogočajo učinkovito analizo velikih zbirk besedil [13].

3.1.1 Predpriprava besedil

S številom dokumentov hitro narašča tudi število različnih besed ali fraz, kar pomeni, da imamo običajno opravka z velikim številom atributov reda 10^5 in več, zato skušamo število uporabljenih besed zmanjšati. V ta namen se poslužujemo postopkov, kot so odstranjevanje nepomembnih besed (*angl. stop-words*) in lematizacija.

Ideja odstranjevanja nepomembnih besed je v odstranitvi besed, ki v sebi nosijo zelo malo informacije. To so predvsem funkcijske besede, med katere spadajo vezniki, zaimki, členki, predlogi in pomožni glagoli. Takšne besede je smiselno izločiti iz proučevanja. Izločanje besed temelji na frekvenci besed, kjer se kot funkcijske besede štejejo besede z visoko frekvenco. Pri tem lahko pride do napak in lahko izločimo tudi besede s pomenom.

Namesto takšnega pristopa se pogosto uporablja že vnaprej pripravljen seznam nepomembnih besed. V našem delu smo seznam nepomembnih besed pripravili sami. Uporabili smo že pripravljene sezname, ki smo jih dopolnili s seznamom najbolj pogostih besed iz našega korpusa spletnih strani, ki je bil predhodno tudi ročno pregledan.

Lematizacija je postopek pretvarjanja besed z lemami oziroma osnovnimi oblikami besed. Lematizacija se pogosto uporablja v rudarjenju besedilnih dokumentov. Na ta način zmanjšamo število različnih besed v korpusu, vendar pa s tem izgubimo dodatne informacije, kot so sklon in slovnični čas. V večini primerov nas zanima le pomen besed, zato je ta postopek zelo razširjen. V našem delu smo uporabili lematizator Lemmagen4J [14], ki je implementacija znanega lematizatorja LemmaGen [15] v programskem jeziku Java.

3.2 Razvrščanje v skupine

Razvrščanje v skupine smo uporabili pri delu s šibko označenimi podatki. S pomočjo teh metod smo iskali skupine šibko označenih podatkov, ki so najbolj podobni ročno označenim podatkom.

3.2.1 Metoda *k* voditeljev

Metoda k voditeljev (*angl. k-means*) je algoritem za razvrščanje podatkov v k skupin. Cilj metode je podane primere razdeliti v k skupin tako, da je vsota razdalj med točkami v skupini in centrom skupine najmanjša [16].

Pri uporabi metode voditeljev moramo že vnaprej poznati število skupin k. Algoritem izbere k voditeljev. Pogosto so voditelji izbrani naključno, obstaja pa vrsta različnih pristopov za izbiro začetnih voditeljev. Splošno pravilo pravi, da naj bodo začetni voditelji med seboj dovolj oddaljeni.

Vsak primer v množici nato priredimo najbližjemu od voditeljev. Primeri, prirejeni istemu voditelju, predstavljajo skupino. Za vsako od skupin izračunamo centroide oziroma središča skupin, ki predstavljajo nove voditelje. Postopek prirejanja primerov najbližjemu od sosedov in izračun centroidov nato ponavljamo, dokler se centroidi ne spreminjajo več.

Mere različnosti Za izračun razdalj med točkami in voditelji lahko izberemo različne mere razdalj. Ker imamo v magistrskem delu opravka s besedilnimi dokumenti, smo izbrali kosinusno podobnost, ki se pogosto uporablja za primerjavo besedil [17].

Kosinusna podobnost je mera podobnosti med dvema vektorjema $\vec{a}$ in $\vec{b}$, ki meri kosinus kota med tema dvema vektorjema. Pomembna je torej usmerjenost vektorjev in ne dolžina. Kosinus kota je lahko izpeljan iz enačbe za skalarni produkt dveh vektorjev:

$$A \cdot B = |A||B| \cos \Theta \tag{3.2}$$

$$podobnost(A, B) = \cos\Theta = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (3.3)$$

Kosinus kota je definiran na območju $[-1, 1]$. Ker pa imamo opravka le s pozitivnimi števili, se njegove vrednosti vedno nahajajo na območju $[0, 1]$. Ker kosinus kota med vektorjema predstavlja podobnost, ga moramo ustrezno spremeniti, da predstavlja razdaljo:

$$razdalja(A, B) = 1 - podobnost(A, B) \quad (3.4)$$

Začetni izbor voditeljev Razvrstitev v skupine, ki je rezultat uporabe metode voditeljev, je lahko zelo odvisna od začetnega izbora voditeljev. Od tega bo tudi odvisno število iteracij, ki privedejo v stabilno stanje. V ta namen obstaja vrsta različnih pristopov; v našem delu smo uporabili implementacijo k-means++ [18], ki je sicer klasična implementacija metode voditeljev, uvaja pa nov način za izbiro začetnih voditeljev, opisan v algoritmu 2.

Algoritem 2: Metoda za razvrščanje v skupine k-means++, ki uvaja nov način za izbiro začetnih voditeljev.

Data: podatki X

Iz množice podatkov X naključno izberi voditelja c_1

for $i \leftarrow 2$ ***to*** k **do**

  Izberi novega voditelja c_i iz $x \in X$ z verjetnostjo $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$, kjer $D(x)$ predstavlja najkrajšo razdaljo med točko in najbližjim že izbranim voditeljem.

end

Nadaljuj s klasičnim algoritmom metode voditeljev.

3.3 Enoznačna klasifikacija

Enoznačna klasifikacija je področje nadzorovanega učenja, pri katerem podatke uvrščamo v natanko enega izmed dveh ali več razredov. Uporabili smo jih kot jedrne metode pri večznačni klasifikaciji, ki delujejo po principu pretvorbe problema večznačne klasifikacije v več problemov enoznačne klasifikacije.

3.3.1 Naivni Bayesov klasifikator

Naloga Bayesovega klasifikatorja [19] je izračunati pogojne verjetnosti za vsak razred pri danih vrednostih atributov za dani novi primer, ki ga želimo klasificirati. Izpeljemo ga s pomočjo Bayesovega pravila:

$$p(c|V) = p(c) * \frac{p(V|c)}{p(V)} \tag{3.5}$$

Naivni Bayesov klasifikator predpostavlja pogojno neodvisnost vrednosti atributov pri danem razredu:

$$p(v_1, v_2, ...v_n|c) = \prod_i p(c|v_i) \tag{3.6}$$

Naivna Bayesova formula:

$$p(c|v_1, v_2, ...v_n) = p(c) * \prod_i \frac{p(c|v_i)}{p(c)} \tag{3.7}$$

Naivni Bayesov klasifikator nov primer klasificira tako, da za vsak možni razred c_i izračuna po naivni Bayesovi formuli izračuna verjetnost, da primer $(v_1, v_2, ...v_n)$ pripada razredu c_i, kar zapišemo $p(c_i|v_1, v_2, ...v_n)$. Primer klasificira v razred z največjo verjetnostjo.

3.3.2 Naključni gozd

Metoda naključni gozd je ansambelska metoda, namenjena izboljšanju napovedne točnosti drevesnih modelov. Ideja je izdelati večje število manjših odločitvenih dreves, tako da se pri izbiri najboljšega atributa v vsakem vozlišču naključno izbere majhno število atributov, ki vstopajo v izbor za najboljši atribut. Običajno se pri m atributih za vsako drevo izbere $\sqrt{m}$ ali $log_2(m)$ atributov. Število zgrajenih dreves je ponavadi 100 ali več.

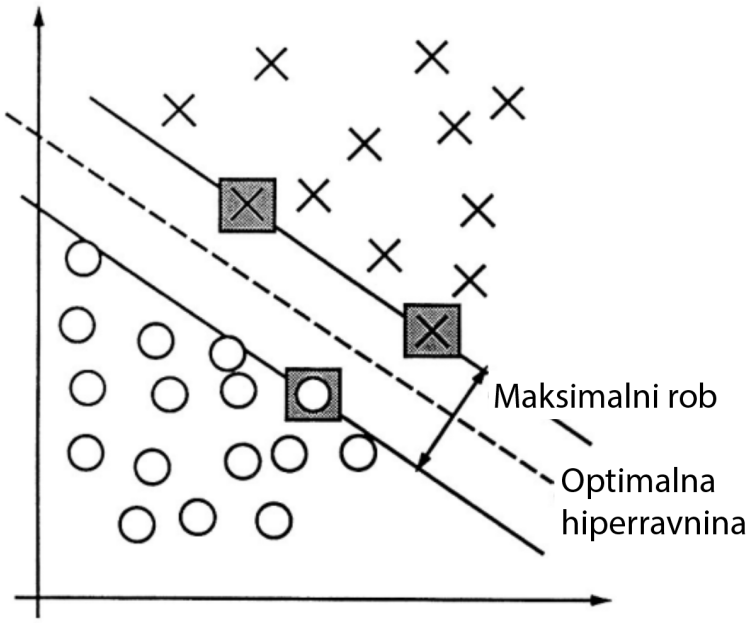
Vsako drevo se uporabi za klasifikacijo novega primera po metodi glasovanja - vsako drevo ima en glas, ki ga nameni razredu, v katerega bi klasificiralo nov primer. Iz vseh glasov dobimo verjetnostno distribucijo po vseh razredih.

Dobra stran naključnih gozdov je običajno visoka točnost, primerljiva z najboljšimi algoritmi, odpornost na šum in osamelce v podatkih [20], slaba stran pa je otežena razlaga odločitve, saj je množica 100 ali več dreves nepregledna in zato nerazumljiva za uporabnika [19].

3.3.3 Metoda podpornih vektorjev

Metode podpornih vektorjev (*angl. Support vector machines - SVM*) so med najbolj uspešnimi metodami za klasifikacijo in regresijo. Primerne so za učenje z velikim številom podatkov, ki so opisani z velikim številom bolj ali manj pomembnih atributov [19].

Osnovna ideja metode je v danem prostoru atributov postaviti optimalno hiperravnino, ki ločuje dva razreda. Optimalna hiperravnina je tista, ki je enako in najbolj oddaljena od najbližjih primerov obeh razredov. Najbližjim primerom optimalne hiperravnine pravimo podporni vektorji, razdalji hiperravnine od podpornih vektorjev pa rob. Torej je optimalna hiperravnina tista, ki ima maksimalni rob [19]. Slika 3.1 prikazuje osnovno idejo metode podpornih vektorjev.



Slika 3.1: Primer hiperravnine z največjim robom do primerov dveh različnih razredov. Primere na teh robovih imenujemo podporni vektorji [21].

3.4 Večznačna klasifikacija

Tradicionalna enoznačna klasifikacija obravnava množico primerov, kjer je vsak primer označen z eno oznako iz množice oznak L. Če je $|L| = 2$, potem govorimo o binarni ali dvorazredni klasifikaciji. Če je $|L| > 2$ gre za večrazredno (*angl. multi-class*) klasifikacijo.

Pri večznačni (*angl. multi-label*) klasifikaciji so primeri povezani z množico oznak $Y \subseteq L$. Takšna klasifikacija se pogosto uporablja v razvrščanju besedil v kategorije. Na primer, članek lahko hkrati pripada več kategorijam, kot so politika, ekonomija, finance [22].

Metode za večznačno klasifikacijo lahko uvrstimo v eno od dveh skupin: a) metode za pretvorbo problema, in b) metode za priredbo algoritmov [22]. Metode iz prve skupine pretvorijo večznačni problem v množico enoznačnih, binarnih problemov, ki se jih nato lotimo z enoznačno klasifikacijo. Metode iz druge skupine pa se ukvarjajo s prilagajanjem algoritmov za enoznačno klasifikacijo, da lahko rešujejo večznačne probleme brez transformacije. Obstaja več priredb algoritmov C4.5 [23], kNN [24] in SVM [25], ki so sposobni večznačne klasifikacije.

V magistrskem delu smo uporabljali predvsem metode za pretvorbo problema, kot so metode binarna relevanca (*angl. binary relevance*), uvrščanje s parno primerjavo (*angl. Ranking by pairwise comparison - RPC*) [26] in ansambelska metoda RAkEL [3].

3.4.1 Binarna relevanca

Binarna relevanca (*angl. Binary relevance*) je ena izmed najbolj preprostih in uporabljenih metod za pretvorbo problema. Po tej metodi se večznačni klasifikacijski problem razdeli na več enoznačnih klasifikacijskih problemov, vsak za eno od oznak v množici oznak $L = y_1, y_2, ..., y_q$.

Metoda najprej pretvori večznačno učno množico v q enoznačnih množic podatkov D_{y_j}, $j = 1 \leq q$, kjer vsaka množica D_{y_j} vsebuje vse primere izvirne množice podatkov, vendar pa pozitivne primere predstavljajo primeri,

ki pripadajo oznaki y_j, negativne pa vsi ostali primeri. Po pretvorbi podatkov se zgradi q enoznačnih klasifikatorjev $H_j(D_{y_j})$, $j = 1 \leq q$, ki so naučeni s pripadajočo množico D_{y_j}. Za večznačno uvrstitev novega primera metoda predlaga vse oznake, katere so enoznačni klasifikatorji napovedali kot pozitiven primer [27].

Pomembna prednost pristopa binarna relevanca je nizka računska zahtevnost v primerjavi z ostalimi metodami za večznačno klasifikacijo. Zahtevnost metode narašča linearno s številom oznak, če pri tem upoštevamo konstantno število primerov. Kljub temu metoda pogosto ni primerna pri velikem številu oznak, saj predpostavlja močno neodvisnost med oznakami in ne upošteva nobene informacije o povezavah med njimi [28].

3.4.2 Uvrščanje s parno primerjavo

Ideja metode RPC (angl. Ranking by pairwise comparison) [26] je transformacija problema s k oznakami $L = \{y_1, y_2, ..., y_k\}$ v $\frac{k(k-1)}{2}$ binarnih problemov, vsakega za en par oznak. Za vsak par oznak $(y_i, y_j) \in L \times L$, kjer velja $1 \leq i < j \leq k$, moramo izdelati klasifikator M_{ij}, ki mora ločiti primere, ki vsebujejo oznako y_i od tistih, ki vsebujejo oznako y_j. Vsak od klasifikatorjev M_{ij} se uči iz primerov, ki vsebujejo eno od teh oznak, vendar pa ne vsebujejo obeh oznak.

Nov primer klasificiramo tako, da ta primer klasificiramo z vsakim od $\frac{k(k-1)}{2}$ klasifikatorjev in nato te napovedi združimo v skupno napoved. Najbolj preprosta metoda za združevanje napovedi je glasovanje vsakega od klasifikatorjev M_{ij} za oznako y_i ali y_j. Na podlagi števila glasov se nato določijo oznake [29].

3.4.3 RAkEL

Nekatere metode za večznačno klasifikacijo (RPC [26], LabelPowerset [22]) imajo slabosti pri velikem številu oznak. Zaradi časovne kompleksnosti takšni problemi postanejo težko rešljivi v razumnem času. Metoda RAkEL je ansambelska metoda, ki omili ta problem, poleg tega pa v mnogih primerih izboljša rezultate [30].

Ideja metode RAkEL (RAndom k labELsets) je, da oznake L naključno razdeli v več manjših skupin oznak $R \subseteq L$ velikosti $k = |R|$. Za vsako od skupin izdela klasifikator z eno od metod za večznačno klasifikacijo. Avtor metode v člankih uporablja večinoma metodo LP, vendar lahko metoda RAkEL deluje praktično s katerokoli metodo za večznačno klasifikacijo [3].

Obstaja več različic algoritma RAkEL, ki se razlikujejo glede na to, kako razdelimo oznake L v manjše skupine. V našem delu smo uporabili algoritem RAkEL$_o$ [3]. Algoritem iz množice vseh kombinacij oznak L velikosti k, ki jo imenujemo L^k, naključno izbere m kombinacij. V primeru, da izberemo število m tako, da velja $mk > |L|$, se bodo te skupine oznak med seboj prekrivale. Nato algoritem začne z učenjem m klasifikatorjev z eno od metod za večznačno klasifikacijo. Postopek učenja metode RAkEL je opisan z algoritmom 3.

Algoritem 3: Učenje metode $RAkEL_o$

Data: množica oznak L velikosti M, učna množica D, velikost podmnožic k, število modelov m

Result: množice k oznak R_i, ustrezni klasifikatorji h_i

S $\leftarrow L^k$;

for $i \leftarrow$ *1* ***to*** *min(m,* $|L^k|$*)* **do**

  $R_i \leftarrow$ naključno izbrana kombinacija oznak iz S;

  učenje klasifikatorja h_i za oznake R_i iz podatkov D;

  S $\leftarrow$ S $\setminus R_i$;

end

Za večznačno klasifikacijo novega, neoznačenega primera se zberejo in združijo napovedi vseh binarnih klasifikatorjev. Združevanje napovedi se naredi z glasovanjem vsakega od klasifikatorjev. Če je delež pozitivnih napovedi za posamezno oznako večji od 0.5, primer označimo s to oznako. Postopek klasifikacije je opisan z algoritmom 4. Tabela 3.1 prikazuje primer klasifikacije primera med 6 različnih oznak $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \}$ z nastavljenimi parametri $k = 3$ in $m = 7$ (m kombinacij oznak velikosti k).

Algoritem 4: Uvrščanje novega primera z metodo RAkEL

Data: množica oznak L velikosti M, število modelov m, kombinacije oznak R_i in njihovi klasifikatorji h_i, nov primer x

Result: *Results* - Vektor oznak za nov primer x

S $\leftarrow L^k$;

for $j \leftarrow$ *1* ***to*** M **do**

  $Sum_j \leftarrow 0$;

  $Votes_j \leftarrow 0$;

end

for $i \leftarrow$ *1* ***to*** m **do**

  forall the *labels* $\theta_j \in R_i$ **do**

    $Sum_j \leftarrow Sum_j + h_i(x, \theta_j)$;

    $Votes_j \leftarrow Votes_j + 1$;

  end

end

for $j \leftarrow$ *1* ***to*** M **do**

  $Avg_j \leftarrow Sum_j / Votes_j$;

  if $Avg_j > 0.5$ **then**

    $Results_j \leftarrow 1$;

  else

    $Results_j \leftarrow 0$;

  end

end

model	množice oznak	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
h_1	$\{\lambda_1, \lambda_2, \lambda_6\}$	1	0	-	-	-	1
h_2	$\{\lambda_2, \lambda_3, \lambda_4\}$	-	1	1	0	-	-
h_3	$\{\lambda_3, \lambda_5, \lambda_6\}$	-	-	0	-	0	1
h_4	$\{\lambda_2, \lambda_4, \lambda_5\}$	-	0	-	0	0	-
h_5	$\{\lambda_1, \lambda_4, \lambda_5\}$	1	-	-	0	1	-
h_6	$\{\lambda_1, \lambda_2, \lambda_3\}$	1	0	1	-	-	-
h_7	$\{\lambda_1, \lambda_4, \lambda_6\}$	0	-	-	1	-	0
	delež glasov	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{2}{3}$
	napoved	1	0	1	0	0	1

Tabela 3.1: Primer klasifikacije novega primera z metodo $RAkEL_o$ s parametri $k = 3$ in $m = 7$ (m kombinacij oznak velikosti k)

Metoda RAkEL torej uspešno zmanjša kompleksnost zahtevnejših metod za večznačno klasifikacijo, pogosto pa tudi izboljša rezultate in zmanjša prekomerno prilagajanje učnim podatkom (*angl. overfitting*). V našem primeru, kjer smo uporabili metodo RPC skupaj z metodo RAkEL, smo kompleksnost zmanjšali iz $\frac{L(L-1)}{2}$ (L - število vseh oznak) na $m \times \frac{k(k-1)}{2}$, kar je precejšnje izboljšanje, saj je parameter k običajno zelo majhno število.

3.5 Mere uspešnosti

Razdelek opisuje mere uspešnosti, ki smo jih uporabili za vrednotenje razvitega postopka za pripravo podatkov občinskega virtualnega asistenta. Mere uspešnosti smo prevzeli iz področja ocenjevanja priporočilnih sistemov in iskanja informacij (*angl. information retrieval*).

3.5.1 Preciznost

Preciznost nam pove, kolikšen delež priporočenih dokumentov je relevantnih. Pri tem smo predpostavili, da za vsako vprašanje in občino predlagamo le en dokument[1]. V našem primeru nam preciznost pove, kolikšen delež predlaganih spletnih strani je relevantnih.

$$Preciznost = \frac{\text{priporočeni relevantni dokumenti}}{\text{priporočeni dokumenti}} \tag{3.8}$$

3.5.2 Srednja recipročna uvrstitev

Srednja recipročna uvrstitev (*angl. Mean Reciprocal Rank*) je statistična mera, ki jo uporabimo takrat, ko za določeno poizvedbo sistem predlaga seznam odgovorov, ki so urejeni po verjetnosti, obstaja pa le en relevanten dokument [4]. Recipročna uvrstitev je recipročna (obratna) vrednost uvrstitve pravilnega odgovora v seznamu predlaganih odgovorov. Srednja recipročna uvrstitev je torej povprečje recipročnih uvrstitev v množici poizvedb Q:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{3.9}$$

Srednja recipročna uvrstitev se pogosto uporablja pri ocenjevanju sistemov za odgovarjanje na vprašanja, oziroma v primerih, ko ima vsaka poizvedba le en pravilen odgovor. V nasprotnem primeru običajno upoštevamo

[1] V sistemu občinskega asistenta se iz tega izbranega dokumenta tvori kratek povzetek strani.

le prvi pravilno predlagan odgovor ali pa izberemo eno od drugih mer za ocenjevanje [31].

Spodnja tabela prikazuje primer izračuna srednje recipročne uvrstitve na primeru sistema, ki besede v ednini prevaja v množino:

Tabela 3.2: Primer izračuna srednje recipročne uvrstitve

poizvedba	predlagani odgovori	pravilen odgovor	uvrstitev	recipročna uvrstitev
roka	**roke**, rok, roki	roke	1	1
otrok	otroki, otroke, **otroci**	otroci	3	1/3
pes	pesi, **psi**, pese	psi	2	1/2
			MRR =	0.61

3.5.3 Priklic@k

Mera priklic@k (*angl. Recall@k*) ocenjuje, kolikšen delež vseh relevantnih dokumentov je bilo uvrščenih med najboljših k predlaganih dokumentov. Priklic@k je definiran kot:

$$R@k(V) = \frac{\sum_{i=1}^{k} rel(v_i)}{\sum_{i=1}^{n} rel(v_i)}, \tag{3.10}$$

kjer je v_i i-ti dokument v množici n dokumentov V in velja $k \leq n$.

V našem primeru imamo za vsako vprašanje in občino le en relevanten dokument, zato nam ta mera v bistvu pove, kolikšna je verjetnost, da se relevanten dokument nahaja med k predlaganimi dokumenti.

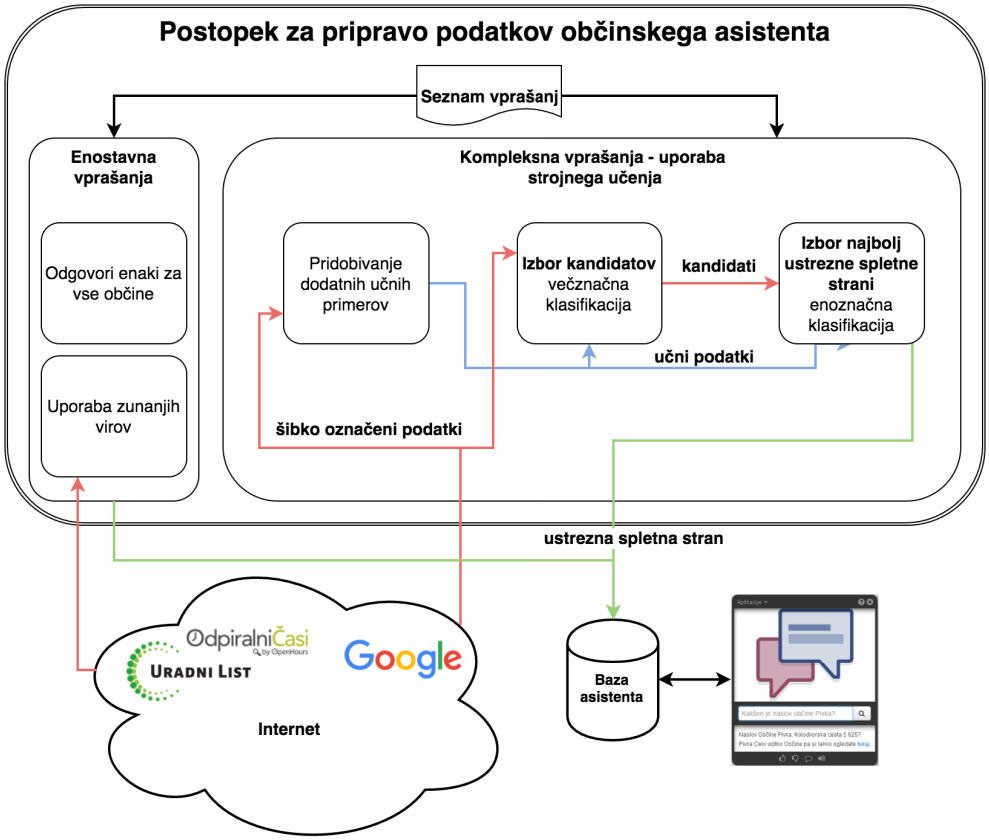
Poglavje 4

Opis postopka priprave podatkov za občinskega asistenta

Postopek za pripravo podatkov za občinskega asistenta je sestavljen iz več podpostopkov, uporabljenih glede na vrsto vprašanj. Kot je že omenjeno v prejšnjih razdelkih, smo vprašanja na grobo razdelili v dve skupini. V prvi skupini so vprašanja, na katere so odgovori enaki vsem občinam ali pa lahko ustrezne strani pridobimo s pomočjo drugih namenskih spletnih aplikacij. V drugi skupini so vprašanja, ki so bolj kompleksna. Za ta vprašanja smo ocenili, da bomo ustrezne spletne strani pridobili s pomočjo strojnega učenja. Slika 4.1 prikazuje shematični prikaz celotnega postopka.

4.1 Preprosti problemi

Ta razdelek opisuje iskanje ustreznih spletnih strani za vprašanja, za katere smo ocenili, da lahko strani najdemo na bolj enostaven način v primerjavi s strojnim učenjem.



Slika 4.1: Shema celotnega postopka za pripravo podatkov za občinskega asistenta.

4.1.1 Enak odgovor za vse občine

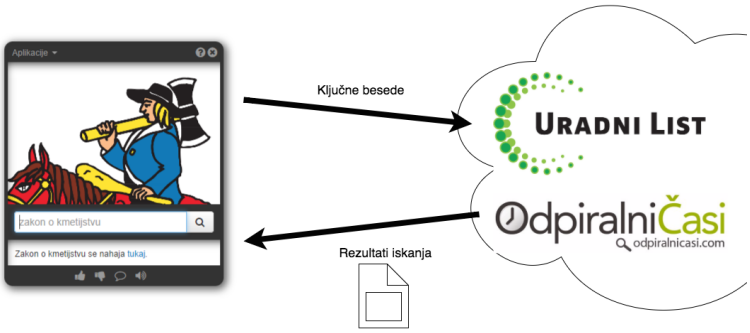
Nekateri odgovori na vprašanja so enaki za vse občine v Sloveniji. Takšna vprašanja se nanašajo predvsem na razne zakone, predpise, obrazce in postopke, ki so enotni v celotni državi. Na takšna vprašanja večinoma odgovarja stran *e-uprava.gov.si*, iz katere smo zato vzeli večino odgovorov tega tipa.

4.1.2 Uporaba zunanjih virov

Nekateri odgovori na vprašanja so specifični za vsako občino, vendar so odgovori enolično določeni. Tovrstne odgovore dobimo s pomočjo uporabe namenskih spletnih strani, ki nam nudijo ustrezen iskalnik in strukturiran prikaz podatkov.

Za pridobitev tovrstnih podatkov smo uporabili dva zunanja vira podatkov. Prvi je spletna stran *uradni-list.si*, kjer so objavljeni zakoni, predpisi in druge javne objave. Z ustreznimi ključnimi besedami in izbiro občine v iskalniku na strani lahko preprosto avtomatsko poiščemo določen dokument. Drugi vir je spletna stran *odpiralnicasi.com*. Ta vir nam omogoča iskanje raznih storitev v bližini posamezne občine (npr. nudi nam seznam kontaktnih informacij o podjetjih in njihovih dejavnostih).

Dobra stran takšnega pristopa je, da so rezultati večinoma pravilni in jih dobimo v strukturirani obliki. Prednost je tudi v tem, da so nekatere spletne strani občin pomanjkljive in ne vsebujejo podatkov o raznih storitvah v občini, zato je uporaba zunanjih virov nujna. Slika 4.2 prikazuje prikaz uporabe zunanjih virov.



Slika 4.2: Uporaba zunanjih virov za iskanje relevantnih spletnih strani.

4.2 Uporaba strojnega učenja

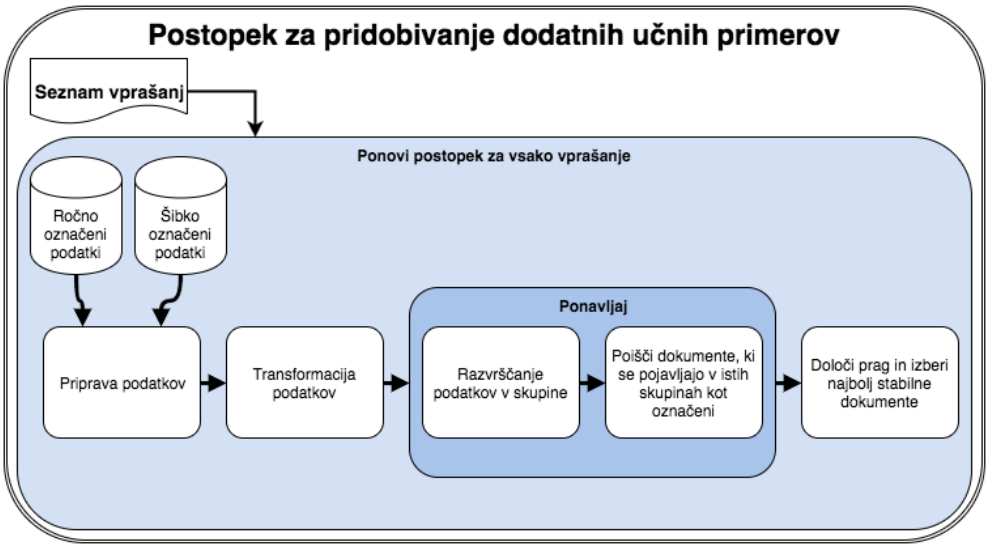
Ta razdelek opisuje postopek, ki z metodami strojnega učenja poišče ustrezne spletne strani za vsako vprašanje. S tem postopkom smo iskali ustrezne spletne strani, ki odgovarjajo na bolj kompleksna vprašanja. Postopek sestavljajo trije deli. Prvi del služi za pridobivanje dodatnih učnih primerov, s katerimi izboljšamo učno množico in posledično napovedi. Drugi del večznačno klasificira spletne strani in predlaga več kandidatov, ki bi lahko predstavljali ustrezen odgovor oziroma ugotavlja, če za določeno vprašanje nimamo nobenega ustreznega odgovora. S tretjim delom tega postopka kandidate razvrstimo po relevantnosti ter izberemo najbolj ustreznega.

4.2.1 Pridobivanje dodatnih učnih primerov

Pogosto nastopi primer, ko je zaradi majhnega števila označenih podatkov učenje klasifikacijskih modelov oteženo, predvsem tam, ko imamo za posamezen razred le nekaj označenih podatkov. Ročno označevanje dodatnih podatkov je namreč časovno zelo potratno, zato potrebujemo avtomatizirano rešitev, ki nam bo našla čim več dobrih učnih primerov.

Ideja postopka za avtomatizirano pridobivanje dodatnih učnih podatkov je, da iz množice novih spletnih strani, ki smo jih pridobili s pomočjo spletnega iskalnika, poiščemo najboljše primere, ki jih bomo dodali učno množico.

Način pridobivanja dodatnih šibko označenih podatkov s spletnim iskalnikom smo že opisali v razdelku 2.2. Postopek za pridobivanje dodatnih učnih primerov iz teh podatkov je prikazan na sliki 4.3 in je bolj podrobno opisan v nadaljevanju.



Slika 4.3: Shematični prikaz postopka za pridobivanje dodatnih učnih primerov.

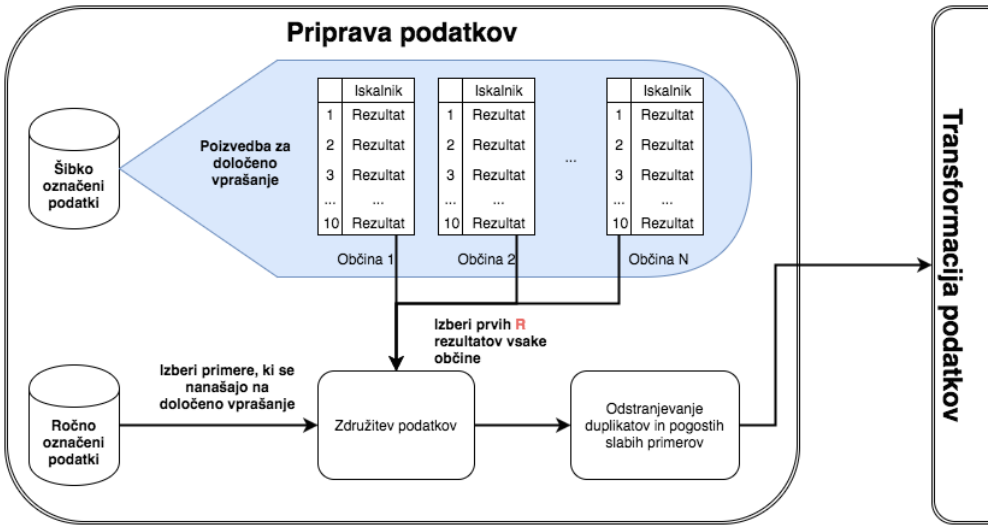
Predlagan postopek za pridobivanje dodatnih učnih primerov sestoji iz sledečih korakov:

Priprava podatkov Celoten postopek pridobivanja novih učnih primerov smo izvedli za vsak razred posebej. Kot vhodne podatke smo uporabili ročno označene in šibko označene podatke. Šibko označeni podatki so v našem primeru spletne strani, ki smo jih pridobili iz spletnega iskalnika in ustrezajo poizvedbi, povezani z vprašanjem ter z občino. Spletni iskalnik za vsako poizvedbo vrne več možnih rezultatov, ki so razvrščeni po relevantnosti.

Tu se pojavi vprašanje, katere dokumente sploh vključiti v postopek. Ali naj uporabimo le dokumente, uvrščene na prvo mesto, ali pa naj uporabimo kar vse dokumente, ki jih iskalnik uvrsti na prvo stran. Izbira vhodnih po-

datkov je torej prvi parameter tega postopka.

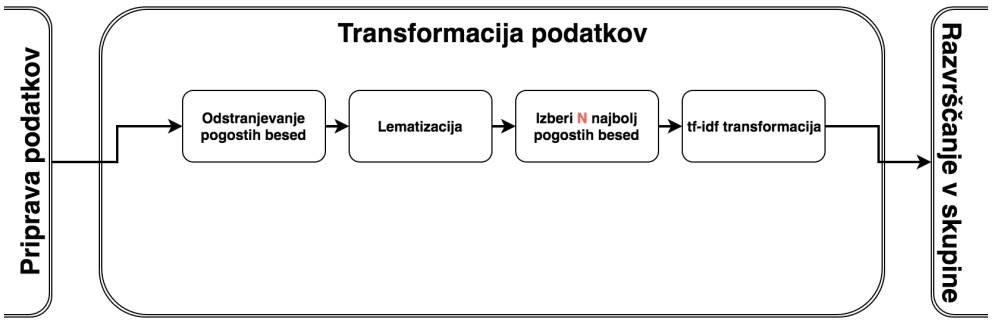
Poleg branja je potrebno tudi primerno čiščenje podatkov. Spletni iskalniki pogosto favorizirajo določene spletne strani, ki se pogosto pojavijo med prvimi zadetki. Po hitrem pregledu vseh podatkov smo izbrali določene spletne strani, ki se pogosto pojavljajo, a jih ne želimo med rezultati (konkretno spletne strani Google Maps, Wikipedia in Uradni list). Vse spletne strani z omenjenimi domenami smo ignorirali in jih nismo upoštevali v nadaljnjih postopkih. Iz podatkov smo odstranili tudi vse duplikate. Slika 4.4 predstavlja shematični prikaz branja in čiščenja podatkov.



Slika 4.4: Shematični prikaz branja in čiščenja podatkov v postopku za pridobivanje dodatnih učnih primerov.

Transformacija podatkov Dokumente je potrebno preoblikovati v obliko, primerno metodi za razvrščanje v skupine. Zato smo besedila spletnih strani lematizirali in odstranili pogoste besede. Seznam pogostih besed smo določili z analizo celotnega nabora pridobljenih spletnih strani in združitvijo že obstoječega seznama pogostih besed [32]. Poseben seznam pogostih besed je potreben zato, da izločimo tudi pogoste besede, ki so specifične za spletne strani občin. V seznam pogostih besed smo dodali tudi imena vseh občin. S tem

smo skušali izničiti vpliv imena občin na podobnost spletnih strani. Besedila smo nato preoblikovali v vektorski prostor s transformacijo *tf-idf*. Elementi vektorja oziroma atributi predstavljajo najbolj pogoste besede. Preizkusili smo več različnih vrednosti števila atributov med 100 in 1000. Slika 4.5 predstavlja shematični prikaz poteka preoblikovanja podatkov v ustrezno obliko.

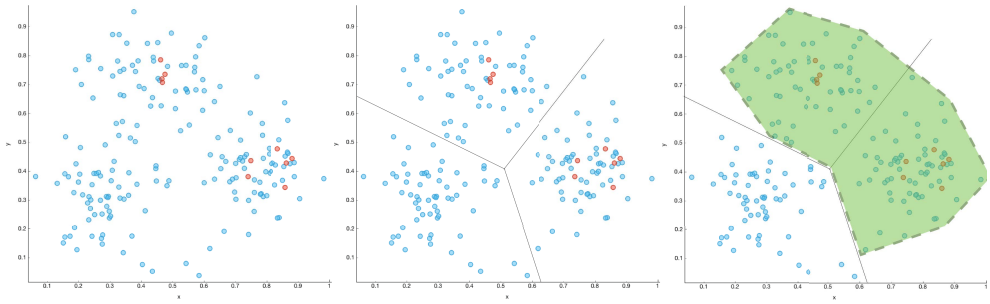


Slika 4.5: Shematični prikaz poteka preoblikovanja podatkov v obliko, ustrezno za nadaljnje razvrščanje v skupine.

Razvrščanje v skupine Ideja tega dela postopka je, da ročno označene podatke in šibko označene podatke, pridobljene s pomočjo spletnega iskalnika, združene razvrstimo v skupine. Predpostavljamo, da elementi iz skupin, ki vsebujejo tudi označene podatke, veljajo za dobre kandidate za vključitev v učno množico 4.6.

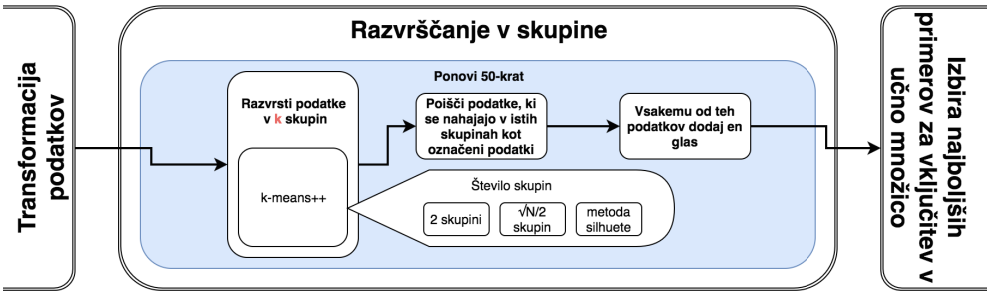
Za razvrščanje v skupine smo uporabili metodo voditeljev, in sicer implementacijo KMeans++, ki bolj pametno razporedi začetne voditelje. Metoda je podrobno opisana v razdelku 3.2.1. Razdalje med dokumenti smo merili s kosinusno razdaljo, ki je primerna za merjenje razdalj med besedilnimi dokumenti. Preizkusili pa smo tri različne načine, na koliko skupin razdeliti podatke: razbitje podatkov na le dve skupini, uporaba pravila palca, ki N podatkov razdeli na $\sqrt{N/2}$ skupin ter z uporabo koeficienta silhuete [33], ki nam pove, na koliko skupin je podatke najbolje razbiti.

Zaradi delno naključne postavitve začetnih voditeljev postopek razvrščanja istih podatkov ne razvrsti vedno v iste skupine. Zato smo celoten postopek



Slika 4.6: Primer odkrivanja skupin na umetnih podatkih. Rdeče točke prikazujejo označene primere, modre točke pa šibko označene primere. Po postopku odkrivanja skupin preverimo, kateri šibko označeni podatki se nahajajo v istih skupinah kot označeni. Ti podatki dobijo glas in predstavljajo kandidate za vključitev v učno množico.

razvrščanja ponovili večkrat, dokler se niso rezultati ustalili. To se je običajno zgodilo pri do 50-kratni ponovitvi razvrščanja. Nato smo opazovali, kateri dokumenti se večinoma pojavljajo v istih skupinah skupaj z označenimi primeri. S tem skušamo doseči, da dodamo v učno množico le tiste dokumente, ki so najbolj stabilni znotraj skupin. Shematični prikaz postopka večkratnega razvrščanja v skupine je prikazan na sliki 4.7.



Slika 4.7: Shematični prikaz večkratnega razvrščanja podatkov v skupine.

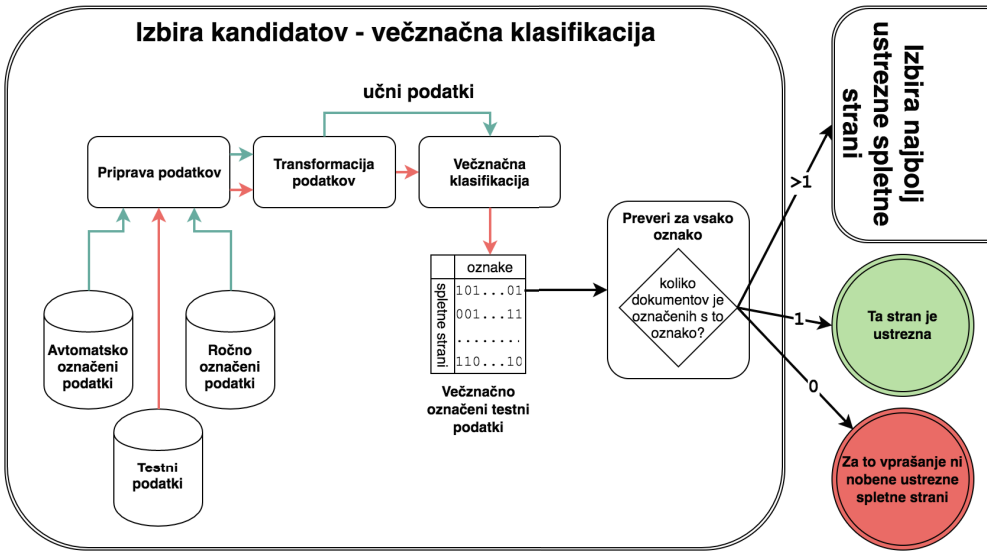
Izbira najboljših primerov za vključitev v učno množico Vprašanje je, katere dokumente nato uvrstiti med učne primere, glede na to, kako pogosto so se ti dokumenti nahajali v istih skupinah kot označeni podatki. Na primer, lahko izberemo tiste, ki so se v istih skupinah kot označeni podatki nahajali v vsaj 50% primerov razvrščanja v skupine. Lahko izberemo na primer najboljših 20% primerov. Skratka, možnosti je veliko. Preizkusili smo več različnih načinov za določanje meje za uvrstitev šibko označenih podatkov v učno množico, opisanih v razdelku 2.2.

4.2.2 Izbira kandidatov z večznačno klasifikacijo

Izbira kandidatov za vsako vprašanje predlaga več spletnih strani, ki bi lahko vsebovale ustrezen odgovor. Ugotovi tudi, če med podatki ni nobene spletne strani, ki bi vsebovala ustrezen odgovor na določeno vprašanje. Posamezna spletna stran lahko odgovarja hkrati na več vprašanj, zato smo se tega problema lotili z večznačno klasifikacijo. Vsako vprašanje smo predstavili s posamezno oznako. Spletne strani se večznačno klasificira in tako ugotovi, na katera vprašanja odgovarjajo. Kot rezultat izbire kandidatov dobimo ponavadi več spletnih strani, ki pripadajo isti oznaki, potrebujemo pa le najbolj ustrezno. Problem izbire najbolj ustrezne spletne strani rešujemo po postopku, opisanem v razdelku 4.2.3. Če določeni oznaki ne pripada noben dokument, sklepamo, da za to vprašanje ni nobene spletne strani z ustreznim odgovorom.

Preizkusili smo dve metodi za večznačno klasifikacijo. Prva je preprosta metoda binarna relevanca, ki za vsako oznako izdela svoj klasifikator ter predpostavlja neodvisnost med oznakami. Druga metoda za večznačno klasifikacijo, ki smo jo preizkusili, je ansambelska metoda RAkEL. V našem primeru metoda RAkEL uporablja množico RPC klasifikatorjev. RPC klasifikatorji transformirajo večznačni problem na več binarnih problemov, vsakega za en par oznak. Torej se osnovni binarni klasifikator uči razlikovati med dvema oznakama.

Postopek izbire kandidatov poteka v treh korakih: izbira podatkov, pred-



Slika 4.8: Shematični prikaz izbire kandidatov z večznačno klasifikacijo

priprava podatkov in večznačna klasifikacija.

Priprava podatkov Za učenje smo uporabili ročno in avtomatsko označene podatke, ki smo jih označili s postopkom za pridobivanje novih učnih primerov.

V primeru uporabe binarne relevance smo za učenje posameznega klasifikatorja uporabili podatke, ki pripadajo določeni oznaki. Iz celotne baze primerov smo naključno izbrali tudi enako število negativnih primerov. To so spletne strani, ki ne pripadajo določeni oznaki. Torej, osnovni binarni klasifikator, naučen s takšnimi podatki, zna oceniti, ali nek nov dokument pripada določeni oznaki ali pa ne.

V primeru uporabe metode RAkEL je osnova množica osnovnih binarnih klasifikatorjev, ki ločujejo med dvema oznakama. Vsak klasifikator torej za učno množico uporablja spletne strani, ki pripadajo eni izmed teh dveh oznak. Uporablja označene spletne strani, kot tudi dodatne spletne strani, avtomatsko označene, pridobljene s postopkom za pridobivanje do-

datnih učnih primerov, opisanim v razdelku 4.2.1. Ker pogosto nimamo na voljo enakega števila primerov za vsak razred, želimo pa enakovrednost vseh vprašanj, smo učno množico uravnotežili s prevzorčenjem.

Transformacija podatkov Spletne strani moramo pretvoriti v obliko, primerno za uporabo različnih metod za klasifikacijo. Besedila spletnih strani smo zato lematizirali in odstranili pogoste besede. Besedilo smo pretvorili v atributno obliko s pomočjo *tf-idf* transformacije in nato izbrali določeno število najboljših atributov.

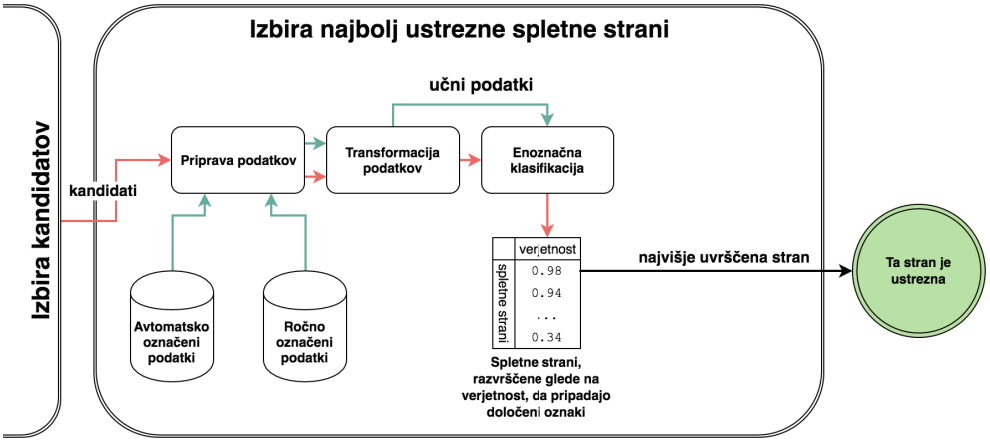
Večznačna klasifikacija Večznačni klasifikator določi, s katerimi oznakami naj označimo določen podatek. Po klasifikaciji izbranih spletnih strani dobimo matriko velikosti $n \times r$ (n - število dokumentov, r - število oznak), iz katere lahko preberemo, katere oznake so bile pripisane posameznim dokumentom. Za vsako od r oznak izberemo pripadajoče spletne strani, ki po klasifikaciji pripadajo tej oznaki, ter jih posredujemo v naslednji korak - izbiro najbolj primerne spletne strani. V primeru, da kateri od r oznak klasifikacija ni pripisala nobene spletne strani, sklepamo, da za vprašanje, povezano s to oznako, ni nobene ustrezne spletne strani.

4.2.3 Izbira najbolj ustrezne spletne strani

Ker nam postopek izbire kandidatov za posamezno oznako navadno predlaga več spletnih strani, moramo med temi predlogi izbrati najbolj ustrezno spletno stran. Metode za večznačno klasifikacijo, ki smo jih uporabili v predhodnem postopku izbire kandidatov, pogosto dajejo le grobe ocene verjetnosti napovedi, zato težko ocenimo, katera spletna stran je najbolj ustrezna za določeno oznako. Zato smo uvedli postopek izbire najbolj ustrezne spletne strani, kjer smo uporabili enoznačno klasifikacijo, ki nam spletne strani enolično razvrsti po njihovi relevantnosti glede na oznako.

Za vsako oznako smo izdelali svoj klasifikacijski model. Vsak tak model zna oceniti, s kakšno verjetnostjo nek dokument pripada določeni oznaki. Tako lahko dokumente razvrstimo po tej verjetnosti in izberemo najbolj ustreznega. Zavedati se moramo, da dobljene verjetnosti običajno niso dobro kalibrirane, so pa uporabne za rangiranje dokumentov [34].

Če pogledamo celoten postopek izbire najbolj ustrezne spletne strani, lahko opazimo, da je postopek skoraj identičen metodi za večznačno klasifikacijo binarna relevanca. V bistvu smo uporabili isto množico klasifikatorjev, tako za metodo binarna relevanca kot tudi za izbiro najbolj ustrezne spletne strani.



Slika 4.9: Izbira najbolj ustrezne spletne strani.

Tudi postopek izbire najbolj ustrezne spletne strani poteka v treh standardnih korakih: priprava podatkov, transformacija podatkov in klasifikacija. Na sliki 4.9 vidimo shematični prikaz postopka izbire najbolj ustrezne spletne strani.

Priprava podatkov Za učenje smo uporabili ročno in avtomatsko označene podatke, ki pripadajo določeni oznaki. Ti podatki predstavljajo pozitivne primere. Iz celotne baze primerov smo naključno izbrali tudi enako število negativnih primerov. Obe podmnožici podatkov potrebujemo v fazi učenja izbranega klasifikatorja, ki prepoznava pripadnost spletne strani določeni oznaki.

Transformacija podatkov Besedila vseh uporabljenih spletnih strani smo lematizirali in iz njih odstranili pogoste besede. Besedila smo pretvorili v atributno obliko s pomočjo transformacije *tf-idf* in nato izbrali določeno število najboljših atributov.

Klasifikacija V tem koraku nam klasifikator za vsako izbrano spletno stran določi verjetnost pripadnosti določeni oznaki. Glede na te verjetnosti smo nato razvrstili spletne strani od najbolj do najmanj ustrezne. Izhajali smo iz dejstva, da lahko te verjetnosti uporabimo za rangiranje rezultatov, če le klasifikatorji vračajo zanesljive ocene verjetnosti [34].

Preizkusili smo več klasifikacijskih metod, za katere je znano, da dajejo zanesljive ocene verjetnosti: metoda podpornih vektorjev, logistična regresija in naključni gozd [35]. Preizkusili smo tudi metodo naivni Bayes, čeprav ocene verjetnosti niso zanesljive, vendar pa so uporabne za rangiranje dokumentov [36].

Poglavje 5

Testna metodologija in rezultati

Celoten postopek priprave podatkov za občinskega asistenta smo ovrednotili na dva načina: avtomatsko in ročno. Avtomatsko vrednotenje rezultatov je primerno, ker je hitro in poceni, vendar pa ima v našem primeru zaradi specifičnosti testnih podatkov nekaj pomanjkljivosti:

- Za določeno vprašanje in občino imamo kot pravilno označeno le eno spletno stran. Seveda pa lahko na isto vprašanje odgovarja več različnih spletnih strani, ki prav tako vsebujejo vse potrebne informacije. Sistem lahko torej predlaga spletno stran, ki je prav tako relevantna vprašanju, vendar bo upoštevana kot napačna.

- Spletne strani upoštevamo le kot pravilne ali nepravilne. Relevantnost ima lahko tudi vmesne vrednosti, s katerimi bi lahko bolje ocenili sistem. Na primer, predlagana je lahko spletna stran, ki je delno relevantna vprašanju, vendar bo upoštevana kot napačna, tako kot če bi bila predlagana popolnoma neustrezna stran.

Kljub pomanjkljivostim nam avtomatska evalvacija omogoča hitro testiranje postopka in dovolj dober vpogled v kvaliteto delovanja. Avtomatsko ocenjevanje smo zato uporabili za optimizacijo parametrov sistema. Da bi dobili tudi realno oceno učinkovitosti postopka priprave podatkov za občinskega asistenta, smo na koncu izvedli tudi ročno evalvacijo.

5.1 Avtomatska evalvacija in izbira parametrov

Postopek smo avtomatsko preverjali s prečnim preverjanjem. Prečno preverjanje na področju večznačne klasifikacije je včasih bolj zahtevno zaradi drugačnih postopkov vzorčenja [37], v našem primeru pa je problem še bolj neobičajen, saj lahko, če gledamo podatke ene občine, posamezna oznaka pripada le enemu primeru. Za testiranje smo simulirali izgradnjo baze za občinskega asistenta.

Za testiranje smo uporabili ročno označene podatke. V vsaki iteraciji smo odstranili podatke ene občine, ki predstavljajo testno množico in ne vpliva na optimizacijo parametrov. Iz preostalih 10 občin smo nato vsakič izbrali tudi podatke ene občine, ki predstavlja validacijsko množico, podatke ostalih občin skupaj z šibko označenimi podatki pa kot učno množico. Za vsako vprašanje smo priporočili 5 spletnih strani iz validacijske množice, urejenih po relevantnosti. Te strani smo primerjali z ročno označeno stranjo za to vprašanje ter izmerili preciznost, srednjo recipročno uvrstitev in priklic@5. Celoten postopek smo ponovili za vse občine tako, da smo vsakič kot validacijsko množico izbrali drugo občino. Na podlagi povprečja rezultatov na validacijski množici smo na koncu izbrali najboljše parametre. Ker je število primerov različno za vsako občino, smo rezultate primerno utežili in izračunali povprečje.

Po izbranih parametrih smo na koncu vsakega od postopkov izvedli tudi preizkus na testnih podatkih po principu izloči enega. To pomeni, da smo vsakič izločili podatke ene občine kot testne podatke, učili pa smo se na ostalih 10 občinah. Rezultati opisanega testa so predstavljeni na koncu vsakega razdelka, kjer je predstavljeno tudi, ali so rezultati po optimizaciji statistično značilni.

5.1.1 Postopek za pridobivanje učnih primerov

Postopek za pridobivanje učnih primerov smo ocenili tako, da smo s tem postopkom najprej izbrali ustrezne učne primere, nato pa izvedli klasifikacijo, ki kot učno množico uporablja tako ročno kot tudi avtomatsko označene primere. Na podlagi metrik za ocenjevanje uspešnosti klasifikacije smo nato ocenili, koliko novi primeri pripomorejo k točnosti klasifikacije. Ta postopek smo ponovili za vsako občino, v vsaki iteraciji smo odstranili podatke tiste občine, na kateri smo testirali postopek (metoda izpusti enega).

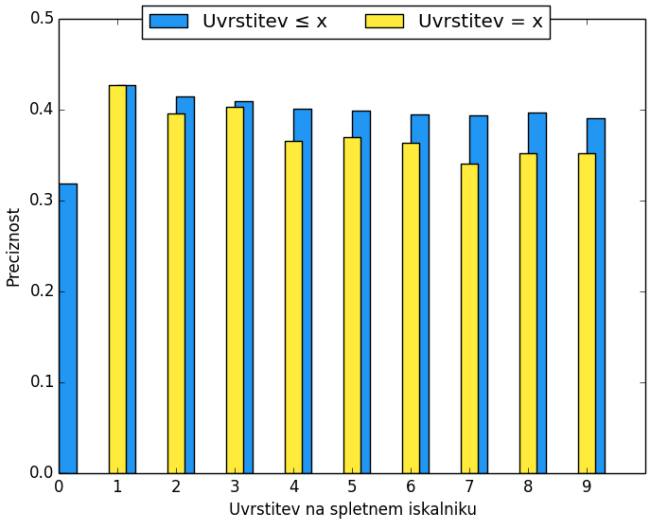
Preizkusili smo štiri različne parametre postopka za pridobivanje učnih primerov in izbrali tiste, s katerimi dobimo najboljše učne primere. Prvi parameter je začetna izbira vhodnih šibko označenih podatkov. Drugi parameter je število atributov, s katerimi predstavimo besedila. Tretji parameter je število skupin, na katere razdelimo podatke z metodo KMeans++ [18]. Zadnji parameter določa prag, katere primere vključimo v učno množico.

Klasifikacijo smo izvedli po postopku, opisanem v razdelku 5.1. Za večznačno klasifikacijo smo uporabili metodo binarna relevanca, pri kateri vsak enoznačni klasifikator uporablja metodo naključni gozd, ki zgradi 100 klasifikacijskih dreves. Besedilo smo transformirali z metodo *tf-idf* in izbrali 100 najbolj pogostih besed med dokumenti, ki pripadajo posamezni oznaki. Postopek klasifikacije je enak v vseh testiranjih postopka za pridobivanje učnih primerov.

Kot merilo za primerjanje smo najprej ocenili, koliko šibko označeni podatki, pridobljeni s pomočjo spletnega iskalnika, pripomorejo k boljši klasifikaciji. Najprej smo ocenili sistem, ki pri učenju ne uporablja šibko označenih podatkov. Nato smo postopoma dodajali šibko označene primere. Najprej smo za učenje dodali primere, ki jih je spletni iskalnik predlagal kot prvi rezultat pri iskanju, nato smo dodali tudi drugo uvrščene in tako dalje. Te meritve služijo kot pregled podatkov in osnovo za primerjavo s postopkom, ki s pomočjo razvrščanja v skupine izbere najbolj primerne podatke za učenje. Cilj je, da bomo z razvrščanjem pridobili boljše učne podatke in posledično dosegli višjo klasifikacijsko točnost.

Slika 5.1 prikazuje preciznost, ki smo jo dobili z uporabo dodatnih šibko označenih podatkov, testirano po principu izpusti eno občino na ročno označenih podatkih. Rumeni stolpci prikazujejo preciznost, ki jo dosežemo, če za učne podatke uporabimo podatke, ki pripadajo pripadajoči uvrstitvi spletnega iskalnika. Modri stolpci pa prikazujejo preciznost, če v učno množico postopoma dodajamo podatke, vsakič z nižjo uvrstitvijo na spletnem iskalniku.

Iz slike je opazno, da šibko označeni podatki močno povišajo kvaliteto klasifikacije. Brez uporabe teh podatkov smo dosegli preciznost 0.32, z uporabo šibko označenih podatkov se preciznost v vseh primerih občutno poviša (nad 0.40). Torej je uporaba teh podatkov smiselna. Iz slike tudi opazimo, da se preciznost znižuje z dodajanjem primerov, ki jih spletni iskalnik uvršča na nižja mesta.



Slika 5.1: Vpliv dodajanja šibko označenih primerov na preciznost.

5.1.1.1 Uvrstitev iskalnika

Zanima nas, katere spletne strani sploh predstavljajo kandidate za vključitev v učno množico. Za vsako vprašanje in občino imamo na voljo prvih deset spletnih strani, ki jih je predlagal spletni iskalnik. Vprašanje je, ali so spletne strani, ki so uvrščene na primer na osmo mesto, sploh še relevantne, ali pa v naše podatke prinašajo preveč šuma?

Celoten postopek pridobivanja učnih podatkov smo ponovili večkrat, v vsaki iteraciji smo kot vhod dodali tudi tiste spletne strani, ki imajo eno stopnjo nižjo uvrstitev. Tabela 5.1 prikazuje rezultate klasifikacije z uporabo učnih primerov, ki smo jih avtomatsko označili. Zanimal nas je torej vpliv uvrstitve na spletnem iskalniku na množico, iz katere pridobivamo nove učne podatke.

Ostale parametre postopka za pridobivanje učnih primerov smo nastavili na privzete vrednosti, ki po ocenah dajejo zadovoljive rezultate. Besedila spletnih strani smo predstavili kot vektorje 200 najbolj pogostih besed v celotni množici. Podatke smo po uveljavljenem palčnem pravilu razdelili na $\sqrt{\frac{N}{2}}$ skupin in na koncu izbrali 50 najbolj stabilnih elementov znotraj skupin, ki vsebujejo označene podatke.

Iz tabele 5.1 vidimo, da vpliv količine podatkov ni velik, če jih le vzamemo dovolj. Če v obdelavo vzamemo vsaj prve tri dokumente vsake občine, se rezultati ustalijo, tudi če dodajamo dokumente z nižjo uvrstitvijo. Kot najboljšo izbiro smo izbrali primer, ko za podatke vzamemo prvih devet dokumentov, ki jih predlaga iskalnik, saj smo s temi podatki dosegli najvišji MRR in priklic@5.

Tabela 5.1: Vpliv izbire vhodnih spletnih strani na kvaliteto pridobivanja novih učnih primerov, glede na uvrstitev spletnega iskalnika.

uvrstitev	Preciznost	MRR	Priklic@5
1	0.408	0.517	0.694
2	0.413	0.531	0.721
3	0.452	0.563	0.743
4	0.452	0.569	0.752
5	0.452	0.564	0.737
6	0.449	0.562	0.746
7	**0.453**	0.567	0.747
8	0.446	0.564	0.743
9	0.451	**0.570**	**0.755**
10	0.439	0.557	0.747

5.1.1.2 Predobdelava besedil

Tabela 5.2 prikazuje rezultate klasifikacije z uporabo učnih primerov, ki smo jih avtomatsko označili. V postopku za pridobivanje učnih primerov smo dokumente predstavili z vektorji, katerih elementi predstavljajo najbolj pogoste besede v celotni množici. Zanima nas, s koliko besedami naj predstavimo spletne strani, da pridobimo čim boljše učne primere.

Tabela 5.2: Vpliv števila atributov, s katerimi predstavimo spletne strani, na kvaliteto pridobivanja novih učnih primerov

število atributov	Preciznost	MRR	Priklic@5
50	0.445	0.561	0.749
100	0.452	0.562	0.739
200	0.451	0.570	**0.755**
300	**0.462**	**0.572**	0.746
400	0.450	0.561	0.736
500	0.461	0.570	0.743
600	0.457	0.569	0.746
700	0.461	0.565	0.735
800	0.457	0.567	0.741
900	0.458	0.567	0.743
1000	0.453	0.561	0.731

5.1.1.3 Število skupin

Zanima nas, na koliko skupin je najbolje razdeliti podatke, da bomo dobili primerno razbitje. Glede na naravo problema bi lahko podatke razbili le na dve skupini in predvidevali, da bodo v eni skupini podatki podobni označenim primerom, v drugi skupini pa večinoma nerelevantni podatki. Vendar tako nastaneta dve veliki skupini in obstaja velika verjetnost, da bodo v skupini z označenimi primeri tudi podatki, ki niso dovolj relevantni.

Druga možnost je razbitje na večje število manjših skupin. Predvidevamo, da bodo primeri znotraj skupin z označenimi primeri v povprečju bolj njim podobni. Vendar pa lahko tako podatke razbijemo na skupine, ki mogoče v resnici niti ne obstajajo.

Na področju odkrivanja skupin, je določanje števila skupin pogost problem. Odločili smo se za tri možnosti. Prva je delitev na dve skupini. Druga možnost je uporaba palčnega pravila, ki predlaga razdelitev N podatkov na

$\sqrt{\frac{N}{2}}$ skupin. Zadnja možnost je avtomatska izbira števila skupin s pomočjo silhuetnega koeficienta.

Tabela 5.3: Rezultati glede na različne pristope za določanje števila skupin

število atributov	Preciznost	MRR	Priklic@5
2 skupini	0.433	0.544	0.725
pravilo palca	**0.462**	**0.572**	**0.746**
silhuetni koeficient	0.454	0.564	0.743

Tabela 5.3 prikazuje rezultate klasifikacije, pri čemer smo v postopku iskanja novih učnih primerov preizkusili različne pristope za določanje števila skupin v množici dokumentov. Najboljše rezultate smo dobili z uporabo palčnega pravila. Tudi uporaba silhuetnega koeficienta daje zadovoljive rezultate, vendar pa smo opazili, da so razlike med koeficienti zelo majhne, zato je z uporabo te metode težko zanesljivo določiti optimalno število skupin.

5.1.1.4 Prag za označitev primera

Po izvedbi prejšnjega koraka – večkratnega razvrščanja elementov v skupine – dobimo seznam spletnih strani in pogostost pripadanja le-teh isti skupini kot označeni podatki. Tu se pojavi vprašanje, kje bomo postavili mejo za dodajanje določenega dokumenta v učno množico.

Preizkusili smo tri različne pristope. Prvi pristop je najbolj preprost in v učno množico sprejme vnaprej določeno število dokumentov (npr. najboljših 50 dokumentov). Drugi pristop upošteva število vseh dokumentov in v učno množico sprejme le določen delež najboljših dokumentov (npr. najboljših 10%). Tretji pristop upošteva le pogostost pojavitev dokumenta v isti skupini kot označeni dokumenti (npr. v učno množico sprejmemo le tiste dokumente, ki so se nahajali v teh skupinah v vsaj 80% primerov).

Tabela 5.4 prikazuje rezultate uporabljenih pristopov. Iz meritev vidimo, da lahko z uporabo kateregakoli pristopa dobimo dobre rezultate, če le na-

stavimo prave vrednosti. Najvišjo preciznost in MRR smo dobili z uporabo tretjega pristopa, kjer smo v učno množico sprejeli le tiste podatke, ki so se v 80% izvajanj postopka odkrivanja skupin nahajali v istih skupinah kot označeni primeri. Ta pristop smo na podlagi merjenj določili za najboljšega. Najvišji priklic@5 smo dobili z uporabo drugega pristopa, vendar pa sta vrednosti preciznost in MRR nižji.

Tabela 5.4: Vpliv različnih pristopov za določanje praga za vključitev primera v učno množico.

	Preciznost	MRR	Priklic@5
Pristop 1			
50	0.462	0.572	0.746
100	0.437	0.556	0.743
Pristop 2			
N/2	0.450	0.561	0.739
N/5	0.458	0.573	**0.766**
N/10	0.447	0.558	0.745
Pristop 3			
0.5	0.455	0.565	0.743
0.7	0.460	0.571	0.750
0.8	**0.466**	**0.576**	0.751
0.9	0.438	0.547	0.716

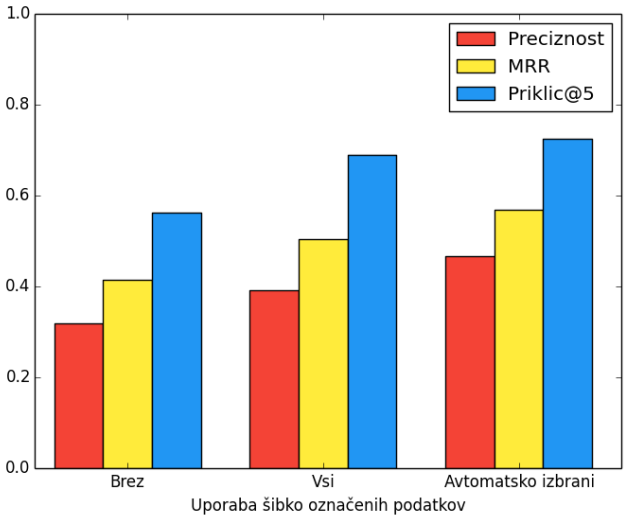
5.1.1.5 Izbrani parametri postopka za pridobivanje dodatnih učnih primerov na podlagi meritev

Na podlagi izvedenih meritev smo izbrali najboljše vrednosti parametrov postopka za pridobivanje novih učnih primerov.

Nove učne primere smo pridobili iz šibko označenih spletnih strani, ki jih je iskalnik uvrstil na eno od prvih devet mest. Postopek smo izvedli za vsako vprašanje posebej. Šibko označene ter ročno označene spletne strani smo združili, jih lematizirali in odstranili pogoste besede, nato pa jih s transformacijo *tf-idf* predstavili z vrečo besed, ki vsebuje 300 najbolj pogostih besed. Vseh N dokumentov smo nato 50-krat razvrstili v $\sqrt{N/2}$ skupin z uporabo metode KMeans++ [18]. V učno množico smo dodali tiste dokumente, ki so se v vsaj 80% primerov pojavili v istih skupinah kot ročno označeni dokumenti.

S tem postopkom smo s preverjanjem na testni množici dosegli preciznost 0.467, MRR 0.568 in priklic@5 0.726. Slika 5.2 prikazuje rezultate v primerjavi klasifikacije brez uporabe šibko označenih podatkov, z uporabo vseh šibko označenih podatkov ter z uporabo šibko označenih podatkov, ki smo jih avtomatsko označili z našim postopkom. Iz slike vidimo, da šibko označeni podatki močno povišajo kvaliteto napovedi. Po izvedbi t-testa smo dobili vrednost $p = 0.000002$, kar zanesljivo potrdi, da je uporaba šibko označenih podatkov smiselna.

Če primerjamo te rezultate s pristopom, ki preprosto uporabi vse šibko označene podatke kot učno množico, smo dosegli izboljšanje 19% v preciznosti, 12% v MRR in 5% v priklic@5. Tudi tu je t-test pokazal, da gre za statistično značilno razliko v rezultatih ($p = 0.038$) in da je uporaba našega pristopa smiselna.



Slika 5.2: Rezultati v primerjavi klasifikacije brez uporabe šibko označenih podatkov, z uporabo vseh šibko označenih podatkov ter z uporabo šibko označenih podatkov, ki smo jih avtomatsko označili z našim postopkom za pridobivanje dodatnih učnih primerov.

5.1.2 Klasifikacija - Binarna relevanca

Ta razdelek opisuje evalvacijo in optimizacijo parametrov za metodo binarna relevanca. Ker ta metoda uporablja iste klasifikatorje kot postopek za izbiro najboljše spletne strani, s tem optimiziramo tudi ta postopek. Celoten postopek evalvacije je opisan v razdelku 5.1.

5.1.2.1 Predobdelava besedil in izbira atributov

Preizkusili smo tri različne metode za izbiro atributov in različna števila izbranih atributov. Metoda, ki smo jo poimenovali max*tf*, je najbolj preprosta in izbere najbolj pogoste besede, za vsak razred posebej. Ocenjevanje atributov z informacijski prispevkom (IG) je zelo pogosto in na splošno daje dobre rezultate, vendar pa obravnava vsak atribut posebej. Mera ReliefF [38] je znana kontekstno-odvisna cenilka pomembnosti atributov, ki zna odkriti včasih zanimive kombinacije atributov.

Tabela 5.5 prikazuje rezultate klasifikacije glede na izbor metode za izbiro atributov. Iz rezultatov je opazno, da vse tri mere dajejo zadovoljive rezultate. Najboljši rezultat smo dobili z najbolj preprosto metodo, ki izbere le najbolj pogoste besede za vsak razred, skupaj 200 atributov.

Tabela 5.5: Rezultati meritev, merjeni z MRR glede na število atributov in metodo za izbiro atributov

število atributov	max*tf*	ReliefF	IG
50	0.563	0.470	0.540
100	0.576	0.501	0.546
200	**0.580**	0. 549	0.558
300	0.578	**0.563**	0.566
400	0.579	0.562	**0.575**
500	0.567	0.556	0.568
750	0.558	0.547	0.555
1000	0.548	0.541	0.539

5.1.2.2 Izbira klasifikatorja in optimizacija parametrov klasifikatorja

Tabela 5.6 prikazuje rezultate klasifikacije glede na izbor metode za klasifikacijo. Primerjali smo štiri metode za klasifikacijo: metoda podpornih vektorjev (SVM), naključni gozd (RF), naivni Bayes (NB) in logistična regresija (LogReg). Iz tabele vidimo, da izbira metode za klasifikacijo močno vpliva na rezultate. Metoda naključni gozd daje občutno boljše rezultate od ostalih metod.

Tabela 5.6: Rezultati klasifikacije glede na izbiro enoznačnega klasifikatorja.

Klasifikator	Preciznost	MRR	Priklic@5
SVM	0.341	0.459	0.654
RF	**0.467**	**0.580**	**0.756**
NB	0.402	0.510	0.682
LogReg	0.410	0.522	0.710

Tabela 5.7 prikazuje rezultate klasifikacije glede na število dreves, ki jih

zgradimo z metodo naključni gozd. S povečevanjem števila dreves kvaliteta klasifikacije narašča, vendar spremembe skoraj niso več vidne, če zgradimo 500 dreves ali več. Najboljši rezultat smo dosegli s 300 drevesi, kjer smo dosegli najvišji MRR in priklic@5.

Tabela 5.7: Rezultati klasifikacije glede na število dreves, ki jih zgradimo z metodo naključni gozd.

Število dreves	Preciznost	MRR	Priklic@5
100	0.467	0.581	0.756
200	0.482	0.592	0.764
300	0.496	**0.600**	**0.771**
400	**0.497**	0.595	0.755
500	0.484	0.590	0.761
600	0.485	0.593	0.761
700	0.485	0.594	0.756
800	0.485	0.593	0.754
900	0.485	0.591	0.757
1000	0.478	0.591	0.760

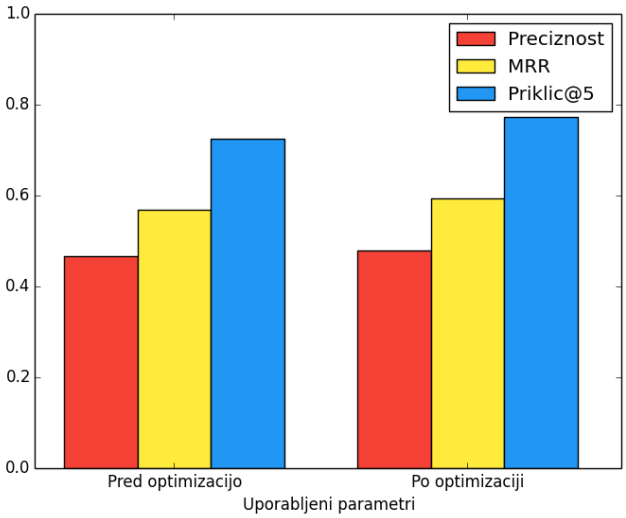
5.1.2.3 Izbrani parametri na podlagi meritev

Na podlagi izvedenih meritev smo izbrali najboljše vrednosti parametrov za postopek klasifikacije z metodo binarna relevanca.

Vsa besedila spletnih strani smo lematizirali in iz njih odstranili najbolj pogoste besede. Nato smo besedila preoblikovali s transformacijo *tf-idf*. Atribute predstavlja 200 najbolj pogostih besed za vsak razred. Vsak enoznačni klasifikator znotraj metode binarna relevanca je metoda naključni gozd, ki zgradi 300 odločitvenih dreves iz naključno izbranih $\lfloor \log_2(N) + 1 \rfloor$ atributov. Enake parametre smo uporabili tudi pri izbiri kandidatov, saj ta postopek uporablja iste enoznačne klasifikatorje kot binarna relevanca.

Z opisanimi parametri klasifikacije smo s testiranjem na testni množici dosegli preciznost 0.479, MRR 0.593 in priklic@5 0.772. Slika 5.3 prikazuje primerjavo rezultatov pred in po optimizaciji parametrov. Z optimizacijo parametrov smo dosegli 3% višjo preciznost, 4% višji MRR in 6% višji priklic@5. Po izvedenem t-testu smo dobili p-vrednost $p = 0.029$, kar pomeni, da je razlika v rezultatih statistično značilna.

0.49585 0.59958 0.77057



Slika 5.3: Rezultati klasifikacije z metodo binarna relevanca pred in po optimizaciji parametrov.

5.1.3 Klasifikacija - RAkEL

Ta razdelek opisuje evalvacijo in optimizacijo parametrov za metodo RAkEL. Postopek evalvacije ostaja enak, kot je opisan v razdelku 5.1.

5.1.3.1 Izbira parametrov metode RAkEL

Delovanje metode RAkEL lahko nastavimo z dvema parametroma k in m (m kombinacij oznak velikosti k), ki sta opisana v razdelku 3.4.3. Preizkusili smo več vrednosti parametra k in povečevali število modelov m, dokler je to še bilo smiselno oziroma postopek večznačne klasifikacije ne traja predolgo. Število modelov m smo predstavili v obliki nM, kjer M predstavlja število vseh oznak, med katere klasificiramo podatke, n pa je poljubno število, s katerim pomnožimo M.

Tabela 5.8: Število atributov

število modelov	Preciznost	MRR	Priklic@k
k = 3			
0.5M	0.368	0.446	0.566
1M	0.459	0.548	0.690
3M	0.473	0.576	0.739
5M	0.475	0.578	0.734
7M	0.475	0.581	0.742
k = 5			
1M	**0.482**	**0.583**	0.740
2M	0.474	0.582	**0.745**
3M	0.463	0.571	0.733

Tabela 5.8 prikazuje rezultate klasifikacije glede na različne parametre metode RAkEL. Uporabili smo tiste parametre za predstavitev podatkov in parametre metod za enoznačno klasifikacijo, ki so dali najboljše rezultate v prejšnjem razdelku 5.1.2.3.

Najboljše rezultate smo dobili pri parametrih $k = 5$ in $m = 3M$. Z opisanimi parametri klasifikacije smo s testiranjem na testni množici dosegli preciznost 0.478, MRR 0.579 in priklic@5 0.732. Rezultati so sicer slabši kot pri uporabi binarne relevance, vendar pričakujemo, da bo metoda RAkEL bolje prepoznavala oznake, za katere ni ustreznih spletnih strani.

5.2 Ročna evalvacija

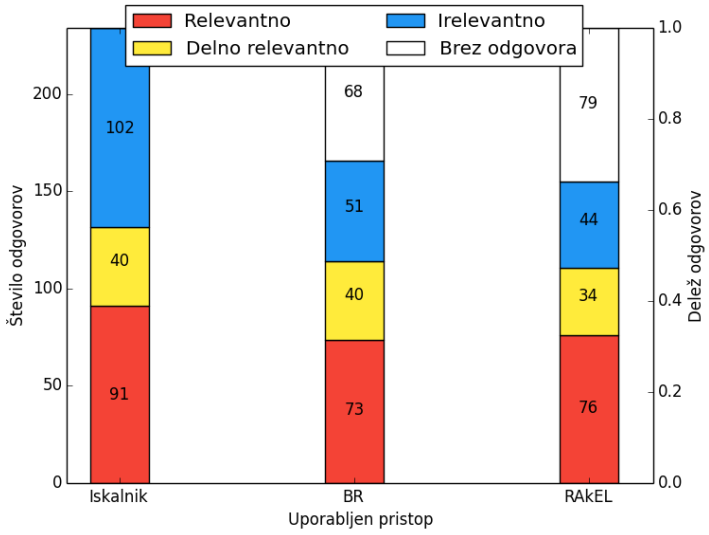
Ročno evalvacijo smo izvedli za 10 naključno izbranih občin. Za vsako od izbranih občin smo izdelali bazo odgovorov. Preverili smo le tista vprašanja, za katere je postopek predlagal spletno stran. Če za določeno vprašanje postopek ni predlagal nobene spletne strani, takšnega rezultata ne moremo preveriti, saj nikoli ne vemo, ali za to vprašanje na spletu res ne obstaja ustrezna spletna stran. Seveda pa nas je zanimal delež neodgovorjenih vprašanj, saj želimo, da je le-ta čim manjši oziroma želimo, da se s povišanjem neodgovorjenih rezultatov zmanjša število neustreznih predlaganih spletnih strani.

Za vsako vprašanje smo za predlagano spletno stran ocenili relevantnost spletne strani glede na vprašanje: relevantno, delno relevantno, irelevantno ali ni odgovora. Relevantno pomeni, da se na spletni strani nahaja iskan podatek. Delno relevantne strani so tiste, na katerih ni iskanega podatka, vendar je tematika predlagane spletne strani podobna tematiki vprašanja. Primer delno relevantne spletne strani je spletna stran občinskega vrtca, ki jo postopek predlaga za vprašanje o vpisu v vrtec. Irelevantne spletne strani so tiste, katerih vsebina nima nobene povezave z vprašanjem. Če postopek ugotovi, da za neko vprašanje nima ustrezne spletne strani, je bolje, da ne priporoči nobene spletne strani, kot pa da nam predlaga irelevantno stran.

Ročno evalvacijo je izvedla kompetentna oseba, zaposlena na Institutu "Jožef Stefan", ki že od začetka celotnega projekta sodeluje pri oblikovanju vprašanj in odgovorov v sodelovanju z občinami.

Ocenili smo oba postopka, razvita za iskanje relevantnih spletnih strani. Prvi za večznačno klasifikacijo uporablja metodo binarna relevanca, drugi pa metodo RAkEL. Kot merilo za primerjavo rezultatov smo ocenili tudi relevantnost spletnih strani, ki jih predlaga spletni iskalnik. Celotni rezultati so predstavljeni v prilogi B.

Slika 5.4 prikazuje povprečne rezultate vseh treh preizkušenih pristopov. Največ relevantnih spletnih strani predlaga spletni iskalnik, vendar ob tem predlaga tudi zelo veliko irelevantnih spletnih strani. Naša razvita pristopa prepoznavata irelevantne odgovore in zato raje ne predlagata nobene spletne



Slika 5.4: Rezultati ročne evalvacije.

strani, kar je pomembna razlika, saj v bazah znanja virtualnih asistentov ne želimo imeti veliko napačnih podatkov. Na ta način v primerjavi s spletnim iskalnikom vsaj razpolovita število predlaganih irelevantnih spletnih strani, z ne prevelikim zmanjšanjem števila relevantnih odgovorov.

Če primerjamo razvita postopka, ki za večznačno klasifikacijo uporabljata metodi binarna relevanca in RAkEL, vidimo, da dajeta podobne rezultate. Metoda RAkEL je na splošno malce bolj stroga in predlaga v povprečju 11 strani manj in tako še malce zmanjša delež irelevantnih odgovorov. Če seštejemo relevantne in delno relevantne strani, sta metodi podobno uspešni.

Poglavje 6

Sklepne ugotovitve

6.1 Izboljšave

Pri razvoju tovrstnih postopkov je vedno prostor za izboljšave. Obstaja še veliko možnosti, ki v magistrskem delu niso omenjene, a so kljub temu vredne nadaljnjega raziskovanja. Nekatere od teh možnosti smo tudi preizkusili, vendar na koncu mnoge od teh niso obetale dobrih rezultatov, zato smo jih tekom dela opustili. Seznam opisuje tiste izboljšave, ki si zaslužijo več pozornosti, vendar niso bile omenjene v tem dokumentu.

Boljša predstavitev spletnih strani V tem delu smo spletne strani obravnavali le kot sezname raznih besed, ki so predstavljale atribute. V tem pogledu bi mogoče lahko marsikaj izboljšali. Spletne strani imajo neko strukturo, na primer naslov, glava, noga in vsebinski del strani. Nekateri deli strani so torej bolj pomembni, drugi manj. Vse to bi lahko upoštevali pri gradnji atributov.

Tekom razvoja smo preizkusili še en pristop k predstavitvi podatkov - analizo arhetipov [39]. Analiza arhetipov je nenadzorovana metoda odkrivanja znanj iz podatkov, ki jo uporabljamo za zmanjšanje dimenzije vhodnih podatkov. Osnovna ideja analize arhetipov je poiskati "čiste tipe" - arhetipe, s katerimi predstavimo vsak podatek v množici kot konveksno kombinacijo množice teh arhetipov.

Metodo smo preizkusili, vendar je zaradi neobetavnih rezultatov in določenih težav uporabljene implementacije na koncu nismo vključili v magistrsko delo. Kljub temu bi bilo mogoče vredno bolj podrobno preizkusiti to metodo.

Uporaba metod delno nadzorovanega učenja Začetni načrt pri izdelavi magistrske naloge je bila uporaba delno nadzorovanega učenja. Nameravali smo uporabiti metodo co-training [40], ki se pogosto uporablja pri klasifikaciji spletnih strani. Co-training temelji na tem, da imamo na podatke dva med seboj neodvisna pogleda. Ideja algoritma je, da s pomočjo obeh pogledov na podatke poskusimo s pomočjo velike in poceni množice neoznačenih podatkov obogatiti majhno množico označenih podatkov.

Uporabo te metode smo opustili zato, ker smo imeli na voljo res premajhno količino podatkov. Za določene oznake, na primer, smo imeli na voljo le en primer, kar je premalo celo za delno nadzorovano učenje. Kot nadomestilo algoritma co-training smo zato razvili postopek za pridobivanje novih učnih primerov, ki deluje na podlagi razvrščanja podatkov v skupine in uporablja šibko označene podatke, ki smo jih pridobili s pomočjo spletnega iskalnika. Na ta način smo dosegli občutno izboljšanje rezultatov klasifikacije.

Ravno na tej točki bi bila dobra uporaba metode co-training, saj imamo končno malce večjo učno množico, ki bi jo še dodatno razširili z uporabo te metode.

Posodabljanje učne množice Pričakujemo, da bodo avtomatsko pripravljeni občinski asistenti spodbudili občine k uporabi in vzdrževanju virtualnih asistentov. S postopkom, predstavljenim v magistrskem delu, smo občutno zmanjšali količino dela, ki ga morajo opraviti zaposleni na občinah za vzdrževanje asistenta.

Postopek za pripravo podatkov občinskega asistenta bi morali izvajati na določen časovni interval, na primer vsak mesec. S tem bomo zagoto-

vili višjo ažurnost podatkov. V prihodnje moramo spremljati tudi vse vnose ter spremembe odgovorov, ki jih vnašajo zaposleni na občini. Ti vnosi bi nam lahko služili za izboljšanje učne množice označenih podatkov. Z večjo količino označenih podatkov pričakujemo, da bo razviti postopek deloval vedno bolje.

6.2 Zaključek

V magistrskem delu smo razvili postopek za pripravo podatkov občinskega virtualnega asistenta. Postopek za različna vprašanja, povezana z dejavnostjo občin, poišče spletne strani, na katerih se nahaja najbolj ustrezen odgovor. Predstavili in uporabili smo različna področja iz strojnega učenja, kot so besedilno rudarjenje, večznačna klasifikacija in uporaba šibko označenih podatkov s pomočjo razvrščanja v skupine. Celoten postopek smo najprej ovrednotili na testnih podatkih, kjer smo prikazali, da uporaba šibko označenih podatkov občutno pripomore k izboljšanju napovedi, poleg tega pa smo prikazali vpliv različnih parametrov postopka na kvaliteto napovedi. Na koncu smo postopek ovrednotili tudi z ročnim pregledovanjem rezultatov. Rezultati so pokazali, da so predlagane spletne strani bolj ustrezne, kot če bi za pripravo podatkov uporabili rezultate spletnega iskalnika. Z razvitimi postopki sicer dobimo malo manjše število relevantnih odgovorov, vendar pa vsaj razpolovimo število irelevantnih odgovorov, kar je zelo pomembno, saj ne želimo baze znanja virtualnih asistentov napolniti z napačnimi podatki. Razviti postopek v povprečju predlaga 76 relevantnih, 35 delno relevantnih in le 44 irelevantnih spletnih strani.

Z razvitim postopkom smo uspešno avtomatizirali postopek priprave podatkov občinskega virtualnega asistenta. S tem smo olajšali delo občinam, ki so odgovore v bazo znanja vnašale ročno, kar je zelo zamudno opravilo. Razviti postopek bo tako pripomogel k hitrejši širitvi projekta Asistent v vse občine.

Dodatek A

Seznam in vpis kategorij

Splošno Vprašanja v zvezi z občino, njeno upravo, občinskim svetom, zaposlenimi, inšpektoratom, zanimivostmi občine, kontaktne informacije občinskih uradnikov (župan, kontakt, uradne ure, občinska uprava, ...)

Vloge in obrazci Iskanje različnih vlog ali obrazcev (vloga za prodajo na sejmu, za izdajo osebnega dokumenta, gradbeno dovoljenje...)

Kultura, šport in izobraževanje Vprašanja, povezana z različnimi športnimi in kulturnimi društvi, izobraževalnimi in kulturnimi ustanovami, prireditvami...

Okolje, prostor in komunala Vprašanja, povezana s prometom, okoljem in komunalo (komunalni prispevek, priključek za vodo, prostorski načrti, razmere na cestah...)

Storitve in obrtniki Vprašanja v zvezi z storitvenimi dejavnostmi v občini (trgovine, bankomati, kinodvorane, banke, obrtniki, ...)

Turizem Vprašanja v zvezi s turizmom (prenočišča, znamenitosti, turistični informacijski center)

Zaščita in reševanje Vprašanja v zvezi s civilno zaščito, gasilci, policijo, načrti zaščite in reševanja

Zdravstvo in sociala Vprašanja v zvezi z zdravstvenim domom, zdravniki, oskrbo na domu, socialno službo...

Kmetijstvo in gospodarstvo Vprašanja v zvezi s kmetijstvom, gozdarstvom, prehrano in gospodarstvom na splošno (predpisi, povračilo škode v primeru suše, prodaja pridelkov, industrija, ...)

Dodatek B

Rezultati ročne evalvacije

Tabela B.1: Rezultati ročne evalvacije - spletni iskalnik

občina	relevantno	delno relevantno	irelevantno	brez odgovora
Kamnik	120	53	62	0
Bled	74	36	125	0
Ankaran	85	33	117	0
Bovec	103	33	99	0
Cerkno	92	55	88	0
Divača	81	34	120	0
Domžale	77	47	111	0
Idrija	83	28	124	0
Izola	102	39	94	0
Grosuplje	95	48	92	0
povprečje	91	41	103	0

Tabela B.2: Rezultati ročne evalvacije - binarna relevanca

občina	relevantno	delno relevantno	irelevantno	brez odgovora
Kamnik	88	64	47	36
Bled	86	39	46	64
Ankaran	69	19	67	80
Bovec	79	23	59	74
Cerkno	73	52	36	74
Divača	82	29	51	73
Domžale	37	55	26	117
Idrija	68	28	73	66
Izola	75	51	49	60
Grosuplje	80	45	60	50
povprečje	74	41	51	69

Tabela B.3: Rezultati ročne evalvacije - RAkEL

občina	relevantno	delno relevantno	irelevantno	brez odgovora
Kamnik	79	54	64	38
Bled	85	29	44	77
Ankaran	74	21	52	88
Bovec	75	25	53	82
Cerkno	76	47	29	83
Divača	83	24	41	87
Domžale	42	43	23	127
Idrija	81	28	48	78
Izola	79	43	37	76
Grosuplje	86	35	49	65
povprečje	76	35	44	80

Dodatek C

Seznam vprašanj

- Napišite, kje najdemo informacije o zaporah na cesti.
- Na kratko predstavite medobčinski inšpektorat in redarstvo.
- Napišite, s pomočjo katere vloge lahko prijavimo prodajo na sejmu.
- Napišite, kako priti do potrdila o oddaljenosti za študente in kam ga poslati.
- Sestavite seznam vlog za najem poslovnega prostora, neprofitnega stanovanja itd.
- Navedite osebo, ki dela v sprejemni pisarni občine, in njene kontaktne podatke.
- Opišite postopek pridobitve zdravstvenega zavarovanja.
- Predstavite referente občinske uprave in podajte njihove kontaktne informacije.
- Na kratko predstavite Svet za preventivo in vzgojo v cestnem prometu.
- Naštejte vse srednje šole v vaši občini in podajte povezave do njihovih spletnih strani.
- Napišite splošen odgovor na vprašanje o kmetijski in gozdarski dejavnosti v vaši občini.
- Najštejte člane volilne komisije občine.
- Ustvarite seznam članov občinskih odborov in komisij ali vstavite povezavo do mesta, kjer so te informacije na voljo.
- Navedite e-poštni naslov občine, namenjen vprašanjem, pripombam in pohvalam občanov oz. obiskovalcev spletne strani občine.
- Naštejte gozdnogospodarske enote v vaši občini.
- Napišite, kako je v vaši občini poskrbljeno za pomoč otrokom v stiski.
- Sestavite seznam zasebnih in splošnih ambulant v občini.
- Opišite gozdarsko dejavnost v vaši občini.
- Naštejte športne oz. večnamenske dvorane v vaši občini in podajte kratek opis ter nekaj koristnih informacij.
- Napišite, kateri zavodi v občini izvajajo program za otroke in mladostnike s posebnimi potrebami.
- Navedite in opišite morebitne muzeje, ki se nahajajo v vaši občini.
- Navedite vse vloge in obrazce v zvezi s kmetijskim zemljiščem.
- Predstavite zgodovinske znamenitosti občine.
- Na kratko predstavite atletski klub (če ga imate) in podajte nekaj koristnih informacij.
- Sestavite seznam vlog, ki zadevajo obratovalni čas gostinskega obrata.
- Predstavite izbrano osnovno šolo in podajte nekaj koristnih informacij v zvezi z njo. Celoten vnos lahko v skrbniških orodjih nato skopirate in vstavite podatke za drugo osnovno šolo.
- Podajte kontaktne informacije društva invalidov v vaši občini.
- Navedite nekaj statističnih podatkov o velikosti občine (površina, dolžina meje).
- Podajte koristne informacije za prepis otroka iz enega vrtca v drugega.
- Predstavite šport v vaši občini.
- Naštejte glasbene skupine oz. ansamble, ki delujejo v vaši občini.
- Navedite, kje lahko najdemo informacije o turizmu v vaši občini.
- Predstavite študentski klub vaše občine.
- Napišite, kdaj praznujete občinski praznik in izbiro datuma razložite.
- Na kratko opišite naloge občinskega sveta in vstavite zanj relevantne povezave.
- Kje najdemo Pravilnik za vrednotenje programov organizacij in društev na področju humanitarnih dejavnosti?
- Opišite postopek pridobitve dovoljenja za gradnjo nezahtevnega objekta.
- Na kratko predstavite župana občine in njegove naloge.
- Napišite, kje lahko najdemo obrazce in vloge v zvezi s komunalo.
- Naštejte vse vloge in obrazce, ki so povezani s prodajo.
- Navedite povezavo do odloka o pomožnih objektih za potrebe občanov in njihovih družin.

- Naštejte/opišite kinodvorane oz. dvorane za javno predvajanje filmov v vaši občini.
- Predstavite športno zvezo vaše občine in športne klube, ki delujejo v njenem okviru.
- Napišite, kje si lahko občani preberejo letni program športa v vaši občini.
- Navedite naslov Turistično informacijskega centra v vaši občini. Navedite tudi mesto, kjer lahko najdemo odlok o ustanovitvi TIC.
- Napišite, kje v občini lahko zainteresirani balinajo.
- Navedite povezavo do Načrta zaščite in reševanja ob jedrski ali radiološki nesreči.
- Na kratko predstavite častne občane občine.
- Navedite aktivnosti, s katerimi se ukvarjajo turistična društva v vaši občini.
- Navedite povezavo do pravilnika o dodeljevanju finančnih sredstev za nakup okolju prijaznih vozil.
- Na kratko opišite posamezno reko, ki teče skozi vašo občino. Če jih je več, tukaj vpišite samo eno in odgovor v urejevalniku samem skopirajte in vnesite potrebne spremembe.
- Navedite povezavo do Sklepa o določitvi in organiziranju enot, služb in drugih operativnih sestavov za zaščito, reševanje in pomoč.
- Napišite vizitko občine (naslov, župan, kontaktne podatke).
- Napišite naslov Zavoda za turizem v vaši občini in dodajte povezavo do odloka o njegovi ustanovitvi.
- Na kratko predstavite odlok o razglasitvi kulturnih in zgodovinskih spomenikov ter naravnih znamenitosti na območju vaše občine ali vstavite povezavo, ki pelje do tega odloka.
- Napišite, kje lahko najdemo javni razpis za sofinanciranje programov društev na področju turizma.
- Opišite postopek za oddajo vloge za finančno pomoč in podajte povezavo na ustrezno vlogo.
- Naštejte krajevne skupnosti občine (če jih imate). V nasprotnem primeru lahko vnos tudi izbrišete.
- Sestavite seznam galerij in razstav v vaši občini in podajte povezave do njihovih spletnih strani ali jih na kratko opišite.
- Naštejte turistične kmetije, ki so na voljo za ogled v vaši občini.
- Navedite vse ponudnike prenočitev v občini.
- Opišite situacijo z rejo drobnice na območju vaše občine.
- Na kratko opišite, kako je v občini urejen avtobusni prevoz šolarjev.
- Predstavite režijski obrat občinske uprave in navedite njegove pristojnosti.
- Navedite povezave do bližnjih gasilskih zvez.
- Opišite postopek za uveljavljanje enkratnega denarnega prispevka za novorojenca.
- Podajte informacije o vpisu otrok v vrtec.
- Napišite, kdaj ima Občina uradne ure.
- Sestavite seznam vseh športnih društev v občini.
- Kje si lahko občani preberejo letni program kulture v vaši občini?
- Podajte osnovne informacije centra za socialno delo.
- Navedite naslov Turistično informacijskega centra v vaši občini.
- Navedite, kje lahko najdemo kinodvorane v vaši občini in dodajte povezave do sporeda, spletnih strani in kontaktnih podatkov.
- Svetujte občanom, kako ravnati v primeru suše.
- Naštejte osnovne šole v vaši občini in podajte povezave do njihovih spletnih strani.
- Navedite informacije o varnostnem sosvetu.
- Kje najdemo sklep o določitvi javnih športnih objektov?
- Sestavite seznam uradnih dokumentov v zvezi s sofinanciranjem obnove in vzdrževanja spomenikov in objektov kulturne dediščine v občini.
- Predstavite kulturni dom v vaši občini in podajte nekaj koristnih informacij v zvezi z njim.
- Predstavite podžupana občine in njegove naloge.
- Napišite, kje lahko najdemo vlogo za pridobitev neprofitnega stanovanja.
- Kje se nahaja Vloga za soglasje na projektne rešitve?
- Sestavite seznam muzejev v vaši občini in podajte povezave do njihovih spletnih strani.
- Navedite povezavo do Načrta zaščite in reševanja ob vremenskih ujmah.
- Navedite povezavo do pravilnika o štipendiranju.
- Sestavite seznam vseh kulturnih društev, klubov in skupin, ki delujejo v občini.
- Navedite povezave do razpisov v letošnjem letu.
- Napišite, kje se nahaja vloga za izdajo soglasja za postavitev pomožnih kmetijsko gozdarskih nezahtevnih in enostavnih objektov.
- Navedite stran, kjer lahko najdemo vlogo za izdajo soglasja za priključitev na cesto oz. javno pot.
- Navedite splošne akte občine.
- Napišite, kje najdemo vlogo za izdajo soglasja o posegu na občinske ceste in javne poti.
- Napišite, kje najdemo vlogo za izdajo potrdila o parcelaciji zemljišča.
- Napišite, kje v vaši občini se nahaja kegljišče in pod kakšnimi pogoji se lahko uporablja.
- Naštejte vse strokovne sodelavce občinske uprave in pripišite njihovo funkcijo.
- Naštejte vse podružnice osnovnih šol in podajte povezave do njihovih spletnih strani (če obstajajo) ali jih na kratko predstavite.
- Vpišite delovni čas medobčinskega inšpektorata.
- Predstavite kulturno dejavnost vaše občine.
- Sestavite seznam večjih podjetij v vaši občini.
- Sestavite seznam sej občinskega sveta ali ustvarite povezavo, ki bo obiskovalce strani pripeljala do njih.
- Naštejte višje in visokošolske ustanove v vaši občini in podajte povezave do njihove spletne strani.
- Navedite naslov občine.
- Podajte kratek opis posameznega vrtca.
- Predstavite vodjo inšpektorata in podajte njegove kontaktne podatke.
- Napišite, kje najdemo obrazec o mesečnem poročilu o številu prenočitev in vplačani turistični taksi.
- Napišite, katere občine obdajajo vašo občino.
- Podajte koristne informacije za odjavo oz. izpis otroka iz izbranega vrtca v občini.
- Na kratko predstavite regionalno televizijsko postajo (če v občini deluje).

- Napišite nekaj o priznanjih, ki jih podeljuje občina (razpis, odlok, kdo jih podeljuje).
- Razložite, kako je urejeno financiranje predšolske vzgoje v vaši občini.
- Navedite spletno stran, kjer so zbrane vse vloge in obrazci.
- Navedite vse svetovalce občinske uprave in njihova delovna področja.
- Napišite, kje so na voljo informacije o prostih delovnih mestih v občini.
- Na kratko opišite zgodovinske znamenitosti občine.
- Na kratko opišite zgodovino vaše občine.
- Navedite odloke v povezavi s čistilno napravo.
- Napišite, kje v vaši občini lahko zainteresirani igrajo badminton.
- Naštejte bazene v občini in aktivnosti, ki se izvajajo v sklopu posameznega bazena. Če bazena v občini nimate, navedite mesto, kjer se lahko občani kljub temu ohladijo oz. kopajo.
- Navedite spletno stran, na kateri so zabeležene prireditve in dogodki v vaši občini.
- Napišite, v katerih ambulantah zdravijo odvisnost od prepovedanih drog.
- Podajte informacije o članih štaba Civilne zaščite in navedite povezavo do Sklepa o imenovanju članov.
- Naštejte vrtce v vaši občini in vstavite povezave do njihovih spletnih strani (če obstajajo) alijih na kratko opišite.
- Navedite informacije o civilni zaščiti in podajte seznam relevantnih odlokov in sklepov.
- Navedite informacije o pokopališčih v vaši občini in dodajte povezavo do sklepa o določitvi cen najema grobnih mest, mrliških vežic in pokopaliških storitev.
- Podajte nekaj zanimivih in koristnih informacij o vojašnici v vaši občini. Če vojašnice v občini nimate, lahko predstavite najbližjo.
- Sestavite seznam zobozdravnikov/zobnih zmbulant v občini.
- Napišite kontaktne podatke občine (telefon, faks).
- Razložite pojem furmanstvo.
- Navedite, kje se nahaja zbirni center komunalnih odpadkov v vaši občini.
- Navedite kontaktne informacije organizacije Rdečega križa v vaši občini.
- Opišite delo občinskih redarjev in njihove pristojnosti.
- Na kratko predstavite pravice in dolžnosti Občinskega sveta.
- Navedite mesto, kjer so shranjeni vsi pomembnejši dokumenti (pravilniki, sklepi, predpisi), ki jih je sprejel Občinski svet.
- Predstavite organizacijo in delovna področja občinske uprave.
- Opišite postopek za oddajo odsluženega vozila ali vstavite povezavo do ustrezne vloge.
- Napišite, kje najdemo vlogo za prodajo blaga s potujočo prodajalno.
- Napišite, kje se lahko v vaši občini igra odbojko.
- Naštejte in na kratko opišite naravne znamenitosti občine.
- Napišite, kje se lahko v vaši občini igra nogomet.
- Opišite industrijsko cono v vaši občini.
- Sestavite seznam vlog v zvezi z zemljišči (splošno).
- Na kratko povzemite vsebino poslovnika Občinskega sveta ali vstavite povezavo do njegove vsebine.
- Podajte kontaktne informacije medobčinskega inšpektorata.
- Podajte odgovor na žalitev župana in občinskih uradnikov.
- Napišite, kje najdemo obrazec za pobudo za spremembo namenske rabe prostora.
- Napišite odgovor na pohvalo občinskih uradnikov.
- Napišite, kje najdemo vlogo za zaporo ceste zaradi prireditve.
- Napišite, katera šola v občini izvaja prilagojen program za otroke z učnimi težavami.
- Naštejte pevske zbore, ki delujejo v vaši občini, in jih na kratko predstavite.
- Naštejte zdravstvene domove v občini in podajte čimveč koristnih informacij v zvezi z njimi.
- Opišite govedorejo v vaši občini.
- Napišite, kje najdemo vlogo za izdajo odločbe o komunalnem prispevku.
- Kje najdemo razpis za zbiranje predlogov za sofinanciranje programov organizacij in društev na področju humanitarnih dejavnosti ter drugih neprofitnih in socialnih dejavnosti?
- Podajte splošen odgovor na vprašanje o cestah.
- Navedite, kje si lahko občani ogledajo aktualne razpise in natečaje ali jih na kratko opišite.
- Napišite, s katero občino je pobratena vaša občina (če je, v nasprotnem primeru lahko vnos izbrišete).
- Podajte informacijo o številu prebivalcev v občini.
- Navedite ime in priimek poslovne sekretarke in njene kontaktne podatke.
- Na kratko predstavite pravice in naloge nadzornega odbora občine in njegove predstavnike.
- Navedite odgovor na splošno vprašanje o občanih.
- Predstavite direktorja občinske uprave in njegove naloge.
- Navedite sedež občinskega davčnega urada in koristne kontaktne informacije.
- Podajte kratek in splošen opis občine.
- Predstavite medobčinski inšpektorat in njegove uslužbence.
- Opišite lego vaše občine.
- Na kratko opišite občinsko glasilo in vstavite povezavo do elektronske oblike glasila (če obstaja).
- Naštejte vsa naselja, ki sestavljajo vašo občino.
- Navedite povezavo do Načrta zaščite in reševanja ob potresu.
- Navedite povezavo do mesta, kjer lahko spremljamo promet v vaši občini.
- Podajte koristne informacije in povezave na vloge za izpis otroka iz vrtca.
- Napišite, kje v vaši občini lahko zainteresirani igrajo namizni tenis.
- Napišite, kje najdemo izjavo najditelja zapuščene živali.
- Navedite znamenitosti vaše občine in naslov Turistično informacijskega centra.
- Navedite, kje lahko najdemo odlok o turistični taksi in morebitne odloke, ki ta odlok dopolnjujejo.
- Naštejte patronažne sestre v občini in podajte nji-

hove kontaktne podatke.
- Napišite, kje lahko najdemo poslovnik nadzornega odbora občine.
- Opišite storitev pomoč družini na domu in njenega izvajalca.
- Navedite povezavo do Načrta zaščite in reševanja ob množičnem pojavu kužnih bolezni.
- Navedite povezavo do seznama gasilskih društev.
- Opišite konjerejo v vaši občini.
- Navedite povezave do načrtov zaščite in reševanja.
- Napišite, kje lahko najdemo različne objave občine (o sklenitvi pogodb, namere, ponudbe itd.).
- Opišite grb občine in razložite njegovo simboliko.
- Navedite povezavo do načrta zaščite in reševanja ob množičnem pojavu kužnih bolezni.
- Navedite, kje lahko najdemo informacije o turističnih znamenitostih občine in kje lahko najdemo turističnega vodiča.
- Navedite vsa društva in zavode, ki nudijo pomoč ob naravnih ali drugih nesrečah.
- Podajte informacije o subvencijah za kmetijstvo.
- Napišite podatke o Zvezi radioamaterjev Slovenije.
- Opišite postopek pridobitve enkratne denarne pomoči za novorojenčka.
- Napišite informacije v zvezi z oskrbo z vodo, vodovodom in komunalo.
- Navedite povezavo do Načrta zaščite in reševanja ob železniški nesreči
- Naštejte naloge poverjenikov za Civilno zaščito in njihovih namestnikov.
- Navedite povezavo do državnega prostorskega načrta za plinovod v vaši občini.
- Sestavite seznam župnijskih karitas v vaši občini.
- Podajte splošen odgovor o gnojenju.
- Naštejte vse vaške skupnosti v vaši občini.
- Sestavite seznam dobrodelnih organizacij, podprtih s strani občine.
- Povejte, kje se nahaja statut občine.
- Navedite volilne enote občine ali povezavo do tega seznama.
- Napišite, kako je v vaši občini urejeno izobraževanje odraslih.
- Napišite, kje se nahaja vloga za prijavo taksnega predmeta oz. uporabo javne površine.
- Napišite, kje lahko najdemo vlogo za (ne)uveljavljanje predkupne pravice.
- Napišite, kje lahko občani najdejo katalog informacij javnega značaja.
- Napišite, kje se lahko v vaši občini igra rokomet.
- Napišite, kje lahko obiskovalci strani najdejo najnovejše novice v zvezi z občino.
- Napišite, s pomočjo katere vloge lahko zainteresirani kupijo stanovanje.
- Opišite dejavnost sadjarstva in živinoreje v vaši občini.
- Predstavite sistemskega administratorja in podajte njegove kontaktne podatke.
- Navedite povezavo do Načrta zaščite in reševanja ob požaru v naravnem okolju
- Opišite pravice in naloge zbora občanov.
- Napišite, kje v vaši občini se nahaja strelišče (če ga imate) in podajte nekaj koristnih informacij v zvezi z njim.
- Podajte nekaj koristnih informacij o smučanju v vaši občini.
- Opišite zastavo občine in razložite njeno simboliko.
- Napišite, kje v vaši občini lahko zainteresirani igrajo košarko.
- Podajte splošen odgovor na vprašanje o klubih, društvih in skupinah.
- Napišite seznam vseh delujočih strank v občini.
- Napišite, kje poteka šola za starše in tečaj opišite.
- Naštejte vloge v zvezi s prostorskim planiranjem in urbanizmom.
- Predstavite knjižnice v vaši občini in podajte nekaj koristnih informacij v zvezi z njimi.
- Napišite, kje lahko najdemo vlogo za izdajo potrdila o namenski rabi zemljišča.
- Navedite, kdo vse opravlja javno gasilsko službo v vaši občini.
- Napišite, kje lahko občani plačajo globo za prekrške, denarne kazni, stroške občinskega organa itd.
- Predstavite glasbeno šolo v vaši občini. Če glasbene šole nimate, povejte, kje se lahko otroci naučijo igrati na različne inštrumente oz. kje lahko obiskujejo tečaj sodobnega plesa.
- Napišite, kje lahko najdemo vlogo za prodajo na stojnici.
- Podajte informacije o dodeljevanju sredstev za subvencioniranje obresti v vaši občini.
- Vnesite povezavo do odloka o rebalansu proračuna občine.
- Napišite, kje lahko občani najdejo vlogo za oprostitev plačevanja storitev oskrbe z vodo in odvajanja odpadnih voda v kmetijstvu.
- Naštejte in na kratko opišite reke, ki tečejo skozi vašo občino.
- Navedite županove svetovalce oz. pomočnike in njihovo funkcijo.
- Na kratko predstavite trško skupnost občine (če jo imate).
- Navedite številko transakcijskega računa občine.
- Navedite odbore in komisije, ki delujejo v sklopu Občine.
- Ustvarite seznam članov občinskega sveta.
- Sestavite seznam zdravnikov, ki opravljajo hišne obiske oz. obiske na domu.

Literatura

[1] Projekt asistent, virtualni asistent za občine in društva. `http://www.projekt-asistent.si/info/`. (Accessed on 02/23/2016).

[2] Leon Noe Jovan, Svetlana Nikić, Damjan Kužnar, and Matjaž Gams. Avtomatizacija izgradnje baze odgovorov virtualnega asistenta za občine in društva. *Inteligentni sistemi: zbornik 17. mednarodne multikonference - IS 2014*, A:46–49, 2014.

[3] G. Tsoumakas, I. Katakis, and L. Vlahavas. Random k-labelsets for multilabel classification. *Knowledge and Data Engineering, IEEE Transactions on*, 23(7):1079–1089, July 2011.

[4] Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan. Evaluating web-based question answering systems. In *In Proc. of LREC 2002*, 2002.

[5] Frank McSherry and Marc Najork. Computing information retrieval performance measures efficiently in the presence of tied scores. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, pages 414–421, Berlin, Heidelberg, 2008. Springer-Verlag.

[6] Alessandro Bergamo and Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 181–189. Curran Associates, Inc., 2010.

[7] Gideon S. Mann and Andrew McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *J. Mach. Learn. Res.*, 11:955–984, March 2010.

[8] Jakramate Bootkrajang and Ata Kabán. Learning a label-noise robust logistic regression: Analysis and experiments. In Hujun Yin, Ke Tang, Yang Gao, Frank Klawonn, Minho Lee, Thomas Weise, Bin Li, and Xin Yao, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, volume 8206 of *Lecture Notes in Computer Science*, pages 569–576. Springer Berlin Heidelberg, 2013.

[9] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Yee-Whye Teh, E. Learned-Miller, and D.A. Forsyth. Names and faces in the news. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–848–II–854 Vol.2, June 2004.

[10] Leon Noe Jovan, Damjan Kužnar, Matjaž Kukar, and Matjaž Gams. Data preparation for municipal virtual assistant. *Inteligentni sistemi: zbornik 18. mednarodne multikonference - IS 2015*, A:47–50, 2015.

[11] Yu-Feng Li, Ivor W. Tsang, James T. Kwok, and Zhi-Hua Zhou. Convex and scalable weakly labeled svms. *J. Mach. Learn. Res.*, 14(1):2151–2188, January 2013.

[12] B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, May 2014.

[13] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 2005.

[14] Slavko Žitnik. Lemmagen4J - Java implementation of Lemmagen tool. `https://github.com/szitnik/Lemmagen4J`. (Obiskano 01.19.2016).

[15] Matjaž Juršič, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. Lemmagen: Multilingual Lemmatisation with Induced Ripple-Down Rules. *Journal of Universal Computer Science*, 16(9):1190–1214, 2010.

[16] Max Bramer. *Principles of Data Mining*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[17] Alexander Strehl, Er Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000*, pages 58–64. AAAI, 2000.

[18] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[19] I. Kononenko and M.R. Šikonja. *Inteligentni sistemi*. Založba FE in FRI, 2010.

[20] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[21] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[22] G. Tsoumakas and I. Katakis. Multi label classification: An overview. *International Journal of Data Warehouse and Mining*, 3(3):1–13, 2007.

[23] Amanda Clare and Ross D. King. *Knowledge Discovery in Multi-label Phenotype Data*, pages 42–53. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[24] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038 – 2048, 2007.

[25] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 14*, pages 681–687. MIT Press, 2001.

[26] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16–17):1897 – 1916, 2008.

[27] Everton Alvares Cherman, Maria Carolina Monard, and Jean Metz. Multi-label problem transformation methods: a case study. *CLEI Electronic Journal*, 14(1):1–10, 2011.

[28] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

[29] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. Technical report, 2010.

[30] Grigorios Tsoumakas and Ioannis Vlahavas. *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings*, chapter Random k-Labelsets: An Ensemble Method for Multilabel Classification, pages 406–417. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[31] Wikipedia. Mean reciprocal rank — wikipedia, the free encyclopedia, 2016. (obiskano 23.2.2016).

[32] Len Dierickx. Github — stopwords, 2015. (obiskano 21.3.2016).

[33] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[34] Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probabi-

lity estimation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2913–2921. Curran Associates, Inc., 2013.

[35] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pages 625–632, New York, NY, USA, 2005. ACM.

[36] Harry Zhang and Jiang Su. *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings*, chapter Naive Bayesian Classifiers for Ranking, pages 501–512. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[37] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6913 of *Lecture Notes in Computer Science*, pages 145–158. Springer Berlin Heidelberg, 2011.

[38] Marko Robnik-Šikonja and Igor Kononenko. An adaptation of Relief for attribute estimation in regression. In Douglas H. Fisher, editor, *Fourteenth International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1997.

[39] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

[40] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM.